Structure-agnostic Causal Representation Learning

Arman Behnam
Department of Computer Science
Illinois Institute of Technology
Chicago, IL, USA
abehnam@hawk.illinoistech.edu

Binghui Wang
Department of Computer Science
Illinois Institute of Technology
Chicago, IL, USA
bwang70@illinoistech.edu

Abstract

Causal representation learning aims to discover robust features by exploiting the causal structure underlying data generation. Existing methods require specifying the causal structure a priori, yet different structures demand fundamentally incompatible invariance constraints, and misspecification leads to representations that discard predictive information. We introduce SaCRL, a framework that jointly identifies which of three canonical causal structures generated the data and learns the corresponding invariant representation without prior structural knowledge. Our approach formulates structure selection as a soft optimization over candidate invariances using HSIC-based violation metrics, with adaptive weights that automatically concentrate on the achievable structure. We provide theoretical guarantees for structure identification, including under random-feature approximation, invariance satisfaction, and out-of-distribution generalization. Empirically, SaCRL recovers the true structure on synthetic and semi-synthetic Bayesian-network benchmarks, outperforms fixed-invariance baselines on Colored MNIST, achieves state-of-the-art accuracy on three DomainBed benchmarks (PACS, VLCS, OfficeHome), and degrades gracefully under structural misspecification and limited environment diversity. Code is available at: https://github.com/ArmanBehnam/sacrl.

1 Introduction

Causal representation learning [39] has emerged as a principled approach to building robust models by uncovering the underlying mechanisms that generate data, enabling generalization beyond training distributions [39, 32]. Current approaches are broadly categorized into two paradigms, each with distinct structural assumptions. **Structure-based** methods [50, 41, 5] require explicit knowledge of Directed Acyclic Graphs (DAGs) and complete Structural Causal Models (SCMs), imposing significant barriers when causal structures are unknown. **Distribution-invariant** methods [4, 21] avoid an explicit graph but enforce a fixed invariance across environments, committing to one causal structure. When that commitment is wrong, the enforced invariance discards predictive signal.

Consider predicting patient outcomes from clinical data collected at several hospitals (Figure 1). Different variables follow different causal patterns: symptoms arise anti-causally from the disease, whose prevalence varies across sites ($\mathcal{G}_1$); biomarkers determine the diagnosis and are themselves shifted by site-specific assay calibration ($\mathcal{G}_2$); and baseline vital signs influence the outcome, which also depends on unobserved, site-dependent severity ($\mathcal{G}_3$). The structure of each variable is unknown to the learner.

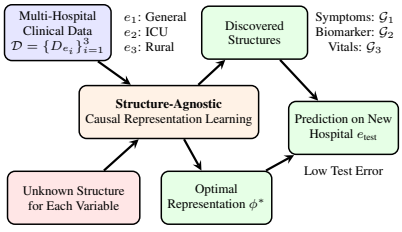


Figure 1: Clinical toy example.

Applying existing methods to such data requires deciding the causal type of each variable a priori. This step is error-prone, because misspecification yields

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

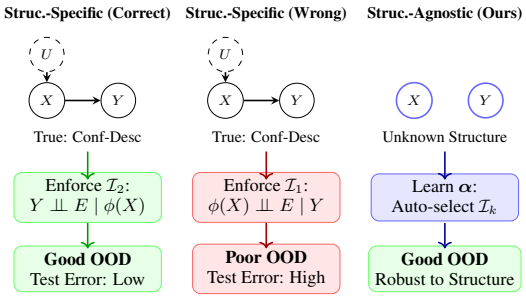
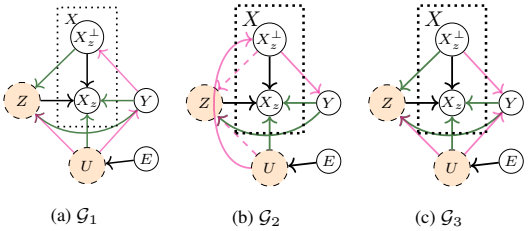


Figure 2: Schematic. **Left**: When the true structure (confounded-descendant, $\mathcal{G}_2$) is known, enforcing its invariance $\mathcal{I}_2$ yields good out-of-distribution performance. **Middle**: Enforcing the invariance of a misspecified structure ($\mathcal{I}_1$) discards predictive information and generalizes poorly. **Right**: SaCRL selects the invariance from data. $\mathcal{I}_k$ denotes the invariance of structure $\mathcal{G}_k$ (Definition 3.1).



(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$ (c) $\mathcal{G}_3$

Figure 3: Three canonical causal structures: (a) anti-causal $\mathcal{G}_1$, (b) confounded-descendant $\mathcal{G}_2$, (c) confounded-outcome $\mathcal{G}_3$. Latent U, Z are shaded, the dotted box is $X = (X_z, X_z^\perp)$, and E acts only through U. Magenta arrows show the edge between $X_z^\perp$ and Y and the confounding by U; the structures differ in the direction of that edge ($Y \to X_z^\perp$ in $\mathcal{G}_1$, $X_z^\perp \to Y$ otherwise) and in the variable that U confounds.

suboptimal representations [54, 23]; combinatorial, because assigning three candidate structures to s groups of variables gives 3^s configurations; and unsupported by guarantees when the structure itself is uncertain [55, 10].

Different causal structures demand incompatible invariance properties [28, 51]. A representation optimized for one structure performs poorly under another, as illustrated in Figure 2. The central question we address is **can we learn optimal invariant representations without knowing the causal structure?** This requires a paradigm shift from *structure-specific* to *structure-agnostic* learning[1]: jointly discover causal mechanisms generating the data and learn representations tailored to those mechanisms. The challenge is both computational (avoiding exponential structure search) and theoretical (providing guarantees when structure itself is uncertain).

We introduce a kernel-based framework, *Structure-agnostic Causal Representation Learning (SaCRL)*, for joint structure discovery and representation learning. SaCRL generalizes Anti-Causal Invariant Abstractions (ACIA) [6], which learns invariant representations for the anti-causal structure $\mathcal{G}_1$, to settings where the structure is unknown. The key theoretical insight is that, when the invariant component of X is predictive, the invariance of the true structure can be satisfied without sacrificing accuracy, whereas the invariances of the other structures incur an irreducible violation. This asymmetry lets a SoftMin objective over the three violations concentrate on the correct structure. We prove that (i) at the minimizers of our objective, the selected structure is correct with probability approaching one exponentially fast in the sample size, provided the penalty weight and the SoftMin temperature are large enough, also under random-feature approximation of the Hilbert–Schmidt independence criterion (HSIC); (ii) the remaining invariance violation is bounded; (iii) the out-of-distribution risk is bounded; and (iv) a sample complexity holds for joint identification and generalization.

2 Background

Random variables are uppercase (X, Y, E), realizations lowercase (x, y, e), and spaces calligraphic ($\mathcal{X}, \mathcal{Y}, \mathcal{E}$). We write $|\mathcal{E}|$ for set cardinality, $\mathbb{E}_P[\cdot]$ for expectation under P, and $\mathbb{I}_A$ for the indicator of set A. Statistical independence is denoted $X \perp\!\!\!\perp Y$ (unconditional) and $X \perp\!\!\!\perp Y \mid Z$ (conditional). Complete notations appear in Appendix B.

2.1 Measure Theory

Following ACIA [6], this work uses measure-theoretic probability. See Appendix C.1 for formal definitions of measurable space, probability measure, product space, σ-algebra, and Markov kernel.

2.2 Canonical Causal Structures

We formalize causal relationships using DAGs in Appendix C.2, which encode the qualitative structure of causal mechanisms. Following Wang and Veitch [51], we consider three canonical causal structures:

[1] It refers to the learner's prior knowledge, not the space of all possible structures. Prior methods require specifying the true structure that governs the data in advance, whereas we eliminate this requirement by learning the structure directly from data.

Definition 2.1. Three Canonical Causal Structures.

(a) **Anti-causal ($\mathcal{G}_1$)**: Observed features decompose as $X = (X_z, X_z^\perp)$ where $Y \to X_z^\perp$, and a latent confounder U influences both Y and the latent variable Z (which in turn affects X_z). DAG includes $Y \to X_z^\perp$ with $U \to Y$ and $U \to Z \to X_z$, capturing settings where labels generate observations.

(b) **Confounded-descendant ($\mathcal{G}_2$)**: Observed features decompose as $X = (X_z, X_z^\perp)$ where $X_z^\perp \to Y$, and a latent confounder U influences both $X_z^\perp$ and other components. DAG includes $X_z^\perp \to Y$ with $U \to X_z^\perp$, and no edge $U \to Y$.

(c) **Confounded-outcome ($\mathcal{G}_3$)**: Observed features decompose as $X = (X_z, X_z^\perp)$ where $X_z^\perp \to Y \to X_z$, and a latent confounder U influences both Y and the latent variable Z (which in turn affects X_z). DAG includes $X_z^\perp \to Y \to X_z$ with $U \to Y$ and $U \to Z \to X_z$.

Here, $X_z^\perp$ denotes the component of X that is causally orthogonal to the Z-mediated pathway. Specifically, X_z captures variation in X that flows through the latent Z, and $X_z^\perp$ captures the remaining variation: a child of Y in $\mathcal{G}_1$, and a parent of Y in $\mathcal{G}_2$ (where it is also affected by U) and in $\mathcal{G}_3$. Figure 3 illustrates the three canonical structures.

The three structures share two conditions: (i) **partial observability:** the learner observes only (X, Y, E), while U and Z are latent; (ii) **environment mechanism:** $E \to U$ is the only edge out of E, so environments shift the distribution of the latent confounder rather than acting on X or Y directly. In all three structures Y affects X_z, and they differ in two places: the direction of the edge between $X_z^\perp$ and Y, and which variable U confounds. In $\mathcal{G}_1$, $Y \to X_z^\perp$ and U confounds Y; in $\mathcal{G}_2$, $X_z^\perp \to Y$ and U confounds $X_z^\perp$; in $\mathcal{G}_3$, $X_z^\perp \to Y$ and U confounds Y. These are the structures compatible with domain shifts that are causally invariant with spurious associations [51]. A bidirectional relation $X \leftrightarrow Y$ violates acyclicity and lies outside the framework. If U has no effect on the variable it confounds, then $(X_z^\perp, Y)$ is independent of E, so all three invariances of Definition 3.1 hold at $X_z^\perp$ and the structures cannot be distinguished.

Definition 2.2 (Environment Space)**.** An environment space consists of a finite collection $\mathcal{E} = \{e_1, \ldots, e_m\}$ of environments that share a measurable input space $(\mathcal{X}, \mathscr{F}_\mathcal{X})$ and a measurable output space $(\mathcal{Y}, \mathscr{F}_\mathcal{Y})$. Each environment $e_i \in \mathcal{E}$ has a probability measure P_{e_i} on the product space $(\mathcal{X} \times \mathcal{Y}, \mathscr{F}_\mathcal{X} \otimes \mathscr{F}_\mathcal{Y})$ and a dataset $D_{e_i} = \{(x_j^{e_i}, y_j^{e_i})\}_{j=1}^{n_{e_i}}$ sampled i.i.d. from P_{e_i}.

Each environment e induces a joint distribution $P_e(X, Y)$ that factorizes according to a causal DAG shared by all environments; only the distribution of the latent confounder U changes with e. We formalize this via causal spaces that pair probability measures with their generating mechanisms (Appendix C.4). Our generalization bound uses a distance $d_\mathcal{K}(e, e')$ between the causal kernels of two environments, which measures how their causal mechanisms differ (Definition D.3).

Definition 2.3 (Representation)**.** A representation is a measurable function $\phi : \mathcal{X} \to \mathcal{Z}$ mapping input space $\mathcal{X}$ to a latent representation space $\mathcal{Z}$, where $(\mathcal{Z}, \mathscr{F}_\mathcal{Z})$ is a measurable space.

3 Problem Formulation

3.1 Motivation

Different causal structures demand fundamentally incompatible invariance properties. An anti-causal setting ($Y \to X$) requires representations conditionally independent of environments given the target. A confounded-descendant setting ($X \to Y$ with confounding) requires the predictive distribution $P(Y|\phi(X))$ to be invariant. A confounded-outcome setting requires the marginal $P(\phi(X))$ to be invariant. A representation optimized for one structure typically performs poorly under another.

Suppose the true structure is confounded-descendant ($\mathcal{G}_2$), but we enforce anti-causal invariance ($\mathcal{I}_1 : \phi(X) \perp\!\!\!\perp E \mid Y$). This constraint forces the representation to discard information that varies with E conditional on Y, but in $\mathcal{G}_2$, this includes precisely the causal signal $X \to Y$. The learned representation loses predictive information. We observe that, when the invariant component $X_z^\perp$ is sufficiently predictive of Y, the correct invariance is achievable (the representation can satisfy it while maintaining prediction accuracy), whereas incorrect invariances are unachievable.

3.2 Problem Setup

Let $\mathcal{E} = \{e_1, \ldots, e_m\}$ with $m \geq 2$ denote the set of environments. For each environment $e \in \mathcal{E}$, we observe a dataset $D_e = \{(x_j^e, y_j^e)\}_{j=1}^{n_e}$ drawn i.i.d. from an unknown joint distribution $P_e(X, Y)$ on the product space $(\mathcal{X} \times \mathcal{Y}, \mathscr{F}_\mathcal{X} \otimes \mathscr{F}_\mathcal{Y})$. The data-generating process follows one of three canonical

causal structures $\mathcal{G}^* \in \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$ shared by all environments, *but the true structure $\mathcal{G}^*$ is unknown to the learner.*

Definition 3.1 (Structure-Specific Invariance)**.** For each causal structure $\mathcal{G}_k$, the corresponding invariance condition $\mathcal{I}_k$ on representation $\phi : \mathcal{X} \to \mathcal{Z}$ is: $\mathcal{I}_1(\phi) : \phi(X) \perp\!\!\!\perp E \mid Y$ for anti-causal, $\mathcal{I}_2(\phi) : Y \perp\!\!\!\perp E \mid \phi(X)$ for confounded-descendant, and $\mathcal{I}_3(\phi) : \phi(X) \perp\!\!\!\perp E$ for confounded-outcome case, where $\perp\!\!\!\perp$ denotes statistical independence and E is the environment indicator.

We formalize the structure-agnostic setting with the product causal space framework. The learner's task is to identify which kernel family governs the observed data. Given data $\mathcal{D} = \{D_{e_i}\}_{i=1}^m$ from m environments generated by an unknown causal structure $\mathcal{G}^* \in \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$, we want to identify the true causal structure $\mathcal{G}^*$ (structure discovery), find $\phi^* : \mathcal{X} \to \mathcal{Z}$ satisfying the invariance $\mathcal{I}_{k^*}$ corresponding to $\mathcal{G}^* = \mathcal{G}_{k^*}$, and learn predictor $w^* : \mathcal{Z} \to \mathcal{Y}$ achieving optimal out-of-distribution generalization simultaneously. Formally, we solve the below objective function:

$$\min_{\phi, w, k} \left\{ \max_{e \in \mathcal{E}} \mathcal{R}_e(w \circ \phi) + \lambda \cdot \mathbb{I}[\neg \mathcal{I}_k(\phi)] \right\} \tag{1}$$

where $\mathcal{R}_e(f) = \mathbb{E}_{(X,Y)\sim P_e}[\ell(f(X), Y)]$ is the risk in environment e. The objective jointly selects the causal structure $\mathcal{G}_k$, learns a representation ϕ satisfying the invariance $\mathcal{I}_k$, and optimizes worst-case risk across environments. Two challenges arise: (i) the structure k is discrete, requiring search over candidates; (ii) the indicator $\mathbb{I}[\cdot]$ is non-differentiable. We address both by observing that each $\mathcal{I}_k$ is an independence statement, which can be quantified continuously via kernel independence testing.

4 Kernel Independence Framework

We develop a kernel-based framework offering rigorous independence characterization and differentiable optimization. Reproducing kernel Hilbert spaces (RKHS) (See Appendix C.5) with characteristic kernels embed probability measures as unique elements.

Definition 4.1 (Characteristic Kernel)**.** A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is characteristic if the kernel mean embedding $\mu : \mathscr{P}(\mathcal{X}) \to \mathcal{H}$ is injective on the space of probability measures $\mathscr{P}(\mathcal{X})$ defined by

$$\mu_P := \mathbb{E}_{X\sim P}[k(\cdot, X)] = \int_{\mathcal{X}} k(\cdot, x)\, dP(x) \tag{2}$$

Common characteristic kernels include the Gaussian RBF kernel $k(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$. For characteristic kernels, $\mu_P = \mu_Q$ if and only if $P = Q$.

We quantify dependence between random variables via the cross-covariance operator $\mathcal{C}_{XY}$, whose Hilbert-Schmidt norm measures dependence strength.

Definition 4.2 (Hilbert-Schmidt Independence Criterion (HSIC))**.** Let $\|\cdot\|_{\mathrm{HS}}$ be the Hilbert-Schmidt norm. The HSIC [13] between random variables X and Y w.r.t. characteristic kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ is:

$$\mathrm{HSIC}(X, Y) := \|\mathcal{C}_{XY}\|^2_{\mathrm{HS}} \tag{3}$$

A key property of HSIC is that it characterizes statistical independence for characteristic kernels.

Lemma 1 (HSIC Characterizes Independence)**.** *Let $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ be characteristic (e.g., Gaussian or delta kernels [44]). Then:*

$$\mathrm{HSIC}(X, Y) = 0 \quad \Longleftrightarrow \quad X \perp\!\!\!\perp Y \tag{4}$$

In practice, we estimate HSIC from finite samples using kernel matrices.

Definition 4.3 (Empirical HSIC)**.** Given samples $\{(x_i, y_i)\}_{i=1}^n$, the empirical HSIC estimator is:

$$\widehat{\mathrm{HSIC}}(X, Y) = \frac{1}{n^2}\mathrm{tr}(K_X H K_Y H) \tag{5}$$

where $[K_X]_{ij} = k_{\mathcal{X}}(x_i, x_j)$, $[K_Y]_{ij} = k_{\mathcal{Y}}(y_i, y_j)$, and $H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ is the centering matrix. To test conditional independence $X \perp\!\!\!\perp Y \mid Z$, where Z is a generic conditioning variable, we employ a residualization approach based on kernel regression.

Definition 4.4 (Kernel Ridge Regression)**.** Let $\{(x_i, z_i)\}_{i=1}^n$ be observations. The kernel ridge regression estimate is:

$$\hat{m}_X(z) := \mathbf{k}_z^\top (K_Z + \gamma I_n)^{-1}\mathbf{x} \tag{6}$$

where $[K_Z]_{ij} = k(z_i, z_j)$, $\mathbf{k}_z = [k(z, z_1), \ldots, k(z, z_n)]^\top$, $\mathbf{x} = [x_1, \ldots, x_n]^\top$, and $\gamma > 0$ is the regularization parameter, chosen via cross-validation or set to $n^{-1/3}$. The residual at observation i is $\tilde{x}_i := x_i - \hat{m}_X(z_i)$.

These residuals remove the dependence of the conditional mean on Z; under the additive-noise model of Lemma 2, this allows us to test conditional independence by applying standard HSIC.

Definition 4.5 (Conditional HSIC)**.** The conditional HSIC for testing $X \perp\!\!\!\perp Y \mid Z$ is:

$$\mathrm{HSIC}(X, Y \mid Z) := \mathrm{HSIC}\big(X - m_X(Z), Y - m_Y(Z)\big) \tag{7}$$

where $m_X(z) = \mathbb{E}[X \mid Z = z]$ and $m_Y(z) = \mathbb{E}[Y \mid Z = z]$. Its empirical estimate is $\widehat{\mathrm{HSIC}}(\tilde{X}, \tilde{Y})$, where $\tilde{X} = X - \hat{m}_X(Z)$ and $\tilde{Y} = Y - \hat{m}_Y(Z)$ are kernel regression residuals.

The lemma below states that kernel regression residuals yield a consistent test for conditional independence under an additive-noise model.

Lemma 2 (Conditional HSIC Consistency)**.** *Assume additive noise, $X = m_X(Z) + \varepsilon_X$ and $Y = m_Y(Z) + \varepsilon_Y$ with $(\varepsilon_X, \varepsilon_Y) \perp\!\!\!\perp Z$, bounded Lipschitz kernels with $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ characteristic, and L^2-consistent regression estimates $\hat{m}_X, \hat{m}_Y$. Then $X \perp\!\!\!\perp Y \mid Z \Leftrightarrow \varepsilon_X \perp\!\!\!\perp \varepsilon_Y$, and:*

1. *If $X \perp\!\!\!\perp Y \mid Z$, then $\widehat{\mathrm{HSIC}}(\tilde{X}, \tilde{Y}) \xrightarrow{p} 0$.*
2. *If $X \not\perp\!\!\!\perp Y \mid Z$, then $\widehat{\mathrm{HSIC}}(\tilde{X}, \tilde{Y}) \xrightarrow{p} \mathrm{HSIC}(\varepsilon_X, \varepsilon_Y) > 0$.*

Under this model, the violation metrics V_1 and V_2, which rely on conditional HSIC, correctly detect whether the corresponding invariances hold. The model does not hold exactly when the regressed variable is discrete, as Y and E are in classification.

5 Structure-agnostic Causal Representation Learning (SaCRL)

Our theoretical findings rely on standard regularity conditions for kernel methods and the representation class, and on the invariant component $X_z^{\perp}$ being sufficiently predictive.

Assumption 1 (Sufficient Environment Diversity)**.** The training environments exhibit variation along three axes: marginal distribution $P(X)$, conditional $P(X|Y)$, and outcome $P(Y|X)$. We assume the three causal structures are distinguishable: testing $\mathcal{I}_1$ requires variation in $P(X|Y)$, testing $\mathcal{I}_2$ requires variation in $P(Y|X)$, and testing $\mathcal{I}_3$ requires variation in $P(X)$.

5.1 Structure Violation Metrics

We quantify the degree to which a representation ϕ violates structure-specific invariance using HSIC.

Definition 5.1 (Population Violation Functions)**.** Let $\phi : \mathcal{X} \to \mathcal{Z}$ be a representation. The population violations are $V_1^*(\phi) := \mathrm{HSIC}(\phi(X), E \mid Y)$ (anti-causal), $V_2^*(\phi) := \mathrm{HSIC}(Y, E \mid \phi(X))$ (confounded-descendant), and $V_3^*(\phi) := \mathrm{HSIC}(\phi(X), E)$ (confounded-outcome).

For empirical violation functions, given pooled observations $\{(x_i, y_i, e_i)\}_{i=1}^N$ from all environments, the empirical violations are $V_1(\phi) := \widehat{\mathrm{HSIC}}(\tilde{\phi}, \tilde{E})$, $V_2(\phi) := \widehat{\mathrm{HSIC}}(\tilde{Y}, \tilde{E}')$, and $V_3(\phi) := \widehat{\mathrm{HSIC}}(\phi(X), E)$. Here, $\tilde{\phi}_i := \phi(x_i) - \hat{m}_\phi(y_i)$ are residuals from regressing $\phi(X)$ on Y, $\tilde{E}_i := e_i - \hat{m}_E(y_i)$ are residuals from regressing E on Y, $\tilde{Y}_i := y_i - \hat{m}_Y(\phi(x_i))$ are residuals from regressing Y on $\phi(X)$, $\tilde{E}'_i := e_i - \hat{m}_{E'}(\phi(x_i))$ are residuals from regressing E on $\phi(X)$, and $\hat{m}(\cdot)$ are kernel ridge regression estimates (Definition 4.4).

The distinction between population (V_k^*) and empirical (V_k) violations is essential. Theoretical results concern V_k^*, while the algorithm optimizes V_k. Consistency ensures $V_k(\phi) \xrightarrow{p} V_k^*(\phi)$ for each fixed ϕ as $n \to \infty$.

5.2 Soft Structure Selection

SoftMin offers a differentiable relaxation for joint structure discovery and representation learning.

Definition 5.2 (SoftMin Aggregation)**.** SoftMin operator on violation metrics with temperature β is:

$$\mathrm{SoftMin}_\beta(\{V_k\}) := -\frac{1}{\beta} \log \sum_k \exp(-\beta V_k) \tag{8}$$

Definition 5.3 (Adaptive Structure Weights)**.** The adaptive weight for structure k given representation ϕ is:

$$\alpha_k(\phi) := \frac{\exp(-\beta V_k(\phi))}{\sum_j \exp(-\beta V_j(\phi))} \tag{9}$$

These weights satisfy $\alpha_k \geq 0$ and $\sum_k \alpha_k = 1$.

Lemma 3 (SoftMin Properties)**.** *SoftMin operator satisfies:*

1. ***Bounds:*** $\min_k V_k - \frac{\log 3}{\beta} \leq \mathrm{SoftMin}_\beta(\{V_k\}) \leq \min_k V_k$
2. ***Limiting behavior:*** $\lim_{\beta\to\infty} \mathrm{SoftMin}_\beta(\{V_k\}) = \min_k V_k$
3. ***Gradient:*** $\nabla_\phi \mathrm{SoftMin}_\beta = \sum_k \alpha_k(\phi)\nabla_\phi V_k$

5.3 Unified Objective Function

Definition 5.4 (SaCRL Objective)**.** The structure-agnostic representation learning objective is:

$$\mathcal{L}(\phi, w) := \underbrace{\max_{e\in\mathcal{E}} \hat{\mathcal{R}}_e(w\circ\phi)}_{\text{worst-case prediction loss}} + \lambda \underbrace{\mathrm{SoftMin}_\beta\left(\{V_k(\phi)\}\right)}_{\text{adaptive invariance penalty}} \tag{10}$$

where $\hat{\mathcal{R}}_e(f) = \frac{1}{n_e}\sum_{j=1}^{n_e} \ell(f(x_j^e), y_j^e)$ is the empirical risk in the environment e.

The objective can be written with explicit structure weights:

$$\mathcal{L}(\phi, w) = \max_{e\in\mathcal{E}} \hat{\mathcal{R}}_e(w\circ\phi) + \lambda \sum_k \alpha_k(\phi) V_k(\phi) - \frac{\lambda}{\beta}\mathrm{Ent}(\boldsymbol{\alpha}(\phi)) \tag{11}$$

where $\alpha_k(\phi)$ are the adaptive weights (Definition 5.3) and $\mathrm{Ent}(\boldsymbol{\alpha}) := -\sum_k \alpha_k \log \alpha_k \leq \log 3$ is their entropy, so the last term vanishes as $\beta \to \infty$.

Definition 5.5 (Distributionally Robust Risk)**.** For risk level $\rho \in (0, 1]$, the conditional value-at-risk (CVaR) over environments is defined as:

$$\mathrm{CVaR}_\rho(\{\hat{\mathcal{R}}_e\}) := \inf_{t\in\mathbb{R}} \left(t + \frac{1}{\rho|\mathcal{E}|}\sum_{e\in\mathcal{E}} \max(\hat{\mathcal{R}}_e(w\circ\phi) - t, 0)\right) \tag{12}$$

CVaR interpolates between the average risk ($\rho = 1$) and the worst-case risk $\max_e \hat{\mathcal{R}}_e$ ($\rho = 1/|\mathcal{E}|$); $\rho = 0.5$ (our default) averages the worst half of the environments, with fractional weight on the boundary environment when $\rho|\mathcal{E}|$ is not an integer.

Definition 5.6 (Distributionally Robust Objective)**.** The distributionally robust structure-agnostic objective is:

$$\mathcal{L}_\rho(\phi, w) := \mathrm{CVaR}_\rho\left(\{\hat{\mathcal{R}}_e(w\circ\phi)\}_{e\in\mathcal{E}}\right) + \lambda\,\mathrm{SoftMin}_\beta\left(\{V_k(\phi)\}_k\right) \tag{13}$$

$\rho = 1/|\mathcal{E}|$ recovers Definition 5.4. We use $\rho = 1/|\mathcal{E}|$ for theoretical analysis and $\rho = 0.5$ in practice.

CVaR_ρ is convex with a closed-form subgradient (Proposition 2, Appendix F).

Optimization: The problem admits a nested decomposition, with the structure weights $\boldsymbol{\alpha}$ in the outer problem and (ϕ, w) in the inner problem (Appendix F.1). Since the outer objective is concave in $\boldsymbol{\alpha}$, its minimizer is a vertex of the simplex (hard selection); SoftMin is its entropy-regularized relaxation, $\mathrm{SoftMin}_\beta(\{V_k\}) = \min_{\boldsymbol{\alpha}\in\Delta^2}\left(\sum_k \alpha_k V_k - \frac{1}{\beta}\mathrm{Ent}(\boldsymbol{\alpha})\right)$, whose minimizer is the adaptive weights $\boldsymbol{\alpha}(\phi)$. The optimization problem is challenging due to (i) the non-convex loss landscape with multiple local minima, and (ii) the difficulty of jointly optimizing all variables. To address these challenges, we adopt the following observations:

1. **Single-loop relaxation:** Instead of solving the nested optimization problem explicitly, we employ a single-loop algorithm in which $\boldsymbol{\alpha}$ is implicitly determined via a SoftMin operator.
2. **Multi-start initialization:** We run the optimization from $M \geq 10$ random initializations and select the solution with the lowest objective value $\mathcal{L}$.
3. **Temperature annealing:** We initialize the SoftMin temperature β with a small value to allow soft selection, and gradually increase it to $\beta_{\max}$ to encourage hard selection.
4. **Hyperparameter selection:** For each λ, multi-start selects the run; across values of λ, we select the model by accuracy on the held-out environment.

Because V_1 regresses the current $\phi(X)$ on Y and V_2 conditions on the current $\phi(X)$, these regressions change during training. Our guarantees are stated at minimizers of the objective (Appendix F.4). In practice, the ERM warm start and temperature annealing stabilize training, and both cross-fitted residuals (Appendix H.8) and computing the regressions on an exponential-moving-average (EMA) encoder (Appendix H.9) give identical results. The soft weights α_k guide optimization: the identified structure is the hard decision $\hat{k} = \arg\min_k V_k(\phi_T)$ made after training (Algorithm 1).

Computation Speedup: Computing empirical HSIC (Definition 4.3) requires $O(n^2)$ time and storage, which is prohibitive for large-scale datasets. To improve scalability, we employ efficient

approximation techniques using the random Fourier features, with Algorithm details in Appendix F. With D random Fourier features ψ [36] (Definition F.1, Appendix F), the approximate HSIC is $\widehat{\mathrm{HSIC}}_{\mathrm{RFF}}(X,Y) = \|\hat{\Sigma}_{\psi_X\psi_Y} - \hat{\mu}_{\psi_X}\hat{\mu}_{\psi_Y}^\top\|_F^2$, which costs $O(nD^2)$ time and $O(nD + D^2)$ memory and avoids the $O(n^2)$ kernel matrices of exact HSIC; memory saving makes kernel penalties feasible on image data (Appendix H.12). The identification guarantee survives this approximation: under the conditions of Theorem D.2, with $D = \Omega(\kappa^4\eta^{-2}d_z\log(n/\delta'))$ random features, where η is the faithfulness gap and d_z the representation dimension, $\hat{k} = k^*$ with probability at least $1 - 6\exp(-n\eta^2/(8192\kappa^8)) - \delta'$ (Corollary 1). *Algorithm details are deferred to Appendix F.*

Selection Confidence: SaCRL reports two confidence measures for $\hat{k}$: the weight margin $\alpha_{(1)} - \alpha_{(2)}$ between the two largest adaptive weights, and the permutation p-value of $\min_k V_k$ against the null threshold τ_α (Definition G.2). A large margin means one structure clearly fits best. If $\min_k V_k > \tau_\alpha$, no candidate invariance holds exactly, and $\hat{k}$ should be read as the closest canonical structure. We calibrate the margin against identification accuracy on data with known structure (Appendix H.6).

Heterogeneous Structures: When different feature subsets follow different causal patterns, we extend the framework to per-feature structure assignment $\sigma^* : \{1,\dots,d\} \to \{1,2,3,\perp\}$ via the objective $\mathcal{L}_{\mathrm{het}}(\phi, w) = \max_e \hat{\mathcal{R}}_e(w\circ\phi) + \lambda\sum_{j=1}^d \mathrm{SoftMin}_\beta(\{V_k^{(j)}(\phi)\}_k)$, applying SoftMin independently per feature. A calibrated permutation threshold $\tau_\alpha^{(j)}$ assigns $\perp$ to features for which no canonical invariance holds: a feature with a canonical structure is wrongly assigned $\perp$ with probability at most α, and a feature with $\sigma^*(j) = \perp$ is assigned $\perp$ with probability tending to one. Per-feature identification guarantees mirror Theorem D.2, and a union bound over the d features adds a $\log d$ term to the sample size. Full treatment is in Appendix G.

6 Theoretical Guarantees

We establish five theoretical results: (1) violation characterization showing the correct structure is identifiable when the invariant component $X_z^\perp$ is sufficiently predictive, (2) finite-sample structure identification holding under random-feature approximation (Corollary 1), (3) a bound on the remaining invariance violation, (4) out-of-distribution generalization bounds, and (5) sample complexity for joint identification and generalization. The details of the theoretical results and proofs are discussed in Appendix D and Appendix E, respectively.

7 Experiments

We evaluate SaCRL across synthetic and real-world settings, addressing four questions: **Q1:** Can the method identify the true causal structure from data? **Q2:** Does correct structure identification improve OOD generalization? **Q3:** Does SaCRL outperform fixed-invariance baselines on standard domain generalization benchmarks? **Q4:** How does it behave under structural misspecification and limited environment diversity? *Additional experiments are deferred to Appendix H.*

7.1 Experimental Setup

Baselines. We compare against: **ERM** (no invariance), **IRM** [4] (enforces $\mathcal{I}_2$, i.e., assumes $\mathcal{G}_2$), **VREx** [21] (variance-based regularization), and **CIRCE** [34] (enforces $\mathcal{I}_1$, i.e., assumes $\mathcal{G}_1$).

Metrics. For synthetic data: we use structure identification accuracy (percentage of runs identifying the correct $\mathcal{G}_k$) and OOD test accuracy on a held-out environment with larger distribution shift. For Colored MNIST, we use OOD accuracy on the anti-correlated test environment, and for DomainBed, accuracy on the held-out domain. Unless stated otherwise, results are means over 5 random seeds, and $\pm$ denotes the standard deviation across seeds.

Implementation Details. For synthetic data, we use 3-layer MLPs with hidden dimensions [256, 128] for both encoder and predictor. For Colored MNIST, we use a 4-layer CNN encoder with base channels 32. For PACS, VLCS, and OfficeHome, we use a frozen ImageNet-pretrained ResNet18 followed by an MLP encoder with hidden dimensions [256, 128] and a linear predictor; the same model produces both the identified structures (Table 4) and the accuracies (Table 5). We select the invariance penalty weight $\lambda \in \{0.5, 1.0, 2.0, 5.0\}$ per experiment by OOD accuracy, i.e., test-domain (oracle) selection in the terminology of Gulrajani and Lopez-Paz [16] (λ=5.0 for Section 7.6; sensitivity in Appendix H.4), and use CVaR with $\rho = 0.5$. We train with Adam optimizer (learning rate 10^{-4} for synthetic, 5×10^{-5} for images) using balanced environment sampling. Random Fourier Features (RFF) with $D = 1000$ dimensions are used to approximate HSIC, reducing memory from $O(n^2)$ to $O(nD + D^2)$ with time linear in n. *Full hyperparameters are provided in Appendix H.1.*

Table 2: Semi-synthetic Bayesian-network benchmarks. Violation: the competing invariance's violation before training, measured in standard deviations of the permutation null. Margin: $\alpha_{(1)} - \alpha_{(2)}$.

Benchmark	Network (nodes)	Y	True	Violation	ID acc. (%)	Margin
asia_g3	ASIA (8)	either	$\mathcal{G}_3$	56.3	95.0 (19/20)	0.45
asia_g2	ASIA (8)	either	$\mathcal{G}_2$	24.6	90.0 (18/20)	0.42
alarm_co	ALARM (37)	CO	$\mathcal{G}_3$	28.3	85.0 (17/20)	0.35
sachs_g2	SACHS (11)	Akt	$\mathcal{G}_2$	6.4	85.0 (17/20)	0.25
alarm_g2	ALARM (37)	EXPCO2	$\mathcal{G}_2$	4.2	75.0 (15/20)	0.18

7.2 Experiment 1: Structure Identification

Setup. We generate synthetic data from each of the three canonical causal structures $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ (Definition 2.1). For each structure, we create $m = 3$ environments with n total samples (divided equally across environments) and $d = 10$ features. We vary the sample size $n \in \{500, 1000, 2000, 5000\}$ to study the effect of data availability on structure identification accuracy. For each $\mathcal{G}_k$, environments differ in noise scale ($\mathcal{G}_1$, $\mathcal{G}_3$) or confounder mean ($\mathcal{G}_2$); weight matrices are sampled from $\mathcal{N}(0, 0.5^2)$ and fixed across runs. Full generative equations are in Appendix H.2.

Results. Table 1 reports structure identification accuracy. Identification accuracy reaches at least 95% at n=5000 and at least 89% at n=2000, demonstrating that the violation-based approach successfully distinguishes the three canonical structures.

Performance improves monotonically with sample size, consistent with the consistency result in Theorem D.2. Structure $\mathcal{G}_3$ is identified most reliably, as its invariance $\phi(X) \perp\!\!\!\perp E$ is unconditional and its violation is estimated without kernel-regression residuals. Structure $\mathcal{G}_2$ is most challenging in small-sample regimes due to the need to estimate conditional independence $Y \perp\!\!\!\perp E \mid \phi(X)$.

Table 1: Structure identification accuracy (%) across sample sizes and true structures. Results averaged over 20 random seeds. Standard deviations in parentheses.

True Structure	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
$\mathcal{G}_1$ (anti-causal)	73.5 (2.8)	85.0 (2.1)	92.5 (1.4)	96.5 (0.9)
$\mathcal{G}_2$ (conf-desc)	68.0 (3.2)	81.5 (2.5)	89.0 (1.8)	95.0 (1.1)
$\mathcal{G}_3$ (conf-out)	78.5 (2.4)	88.5 (1.9)	94.0 (1.2)	97.5 (0.7)

Semi-synthetic Bayesian networks. To test identification where the ground truth is fixed independently of our method, we build five benchmarks from three published Bayesian networks (ASIA, SACHS, ALARM). For each, we choose a target Y, take X as its Markov blanket, induce environments by soft interventions on a latent parent outside the blanket, and read the canonical structure off the published DAG by the rules of Section 2; each structure is verified in simulation before training (Appendix H.7). These topologies admit $\mathcal{G}_2$ and $\mathcal{G}_3$ but not $\mathcal{G}_1$, which would require the Markov blanket of Y to consist only of its descendants. SaCRL selects the correct structure in the majority of runs (75–95%, Table 2), whereas returning any single structure by default is correct on at most 3/5. Accuracy and margin are highest on the most separated benchmark (asia_g3) and lowest on the least separated one (alarm_g2), consistent with the dependence of Theorem D.2 on the faithfulness gap η, for which the pre-training violation in Table 2 is a rough proxy.

7.3 Experiment 2: OOD Generalization

Setup. Following Arjovsky et al. [4], we construct a binary classification task on MNIST digits (0–4 vs. 5–9) where color is spuriously correlated with the label. Specifically, for each digit, we assign color (red or green) based on its label with probability ρ_e in environment e. Training uses two environments with $\rho_1 = 0.9$ and $\rho_2 = 0.8$, while the test environment has $\rho_{\text{test}} = 0.1$ (anti-correlation). The label causes the color assignment, so we treat the anti-causal $\mathcal{G}_1$ as the reference structure. Since the environment changes the color mechanism, the digit shape satisfies all invariances.

Results. We subsample 5000 examples per training environment and 5000 test examples to ensure balanced evaluation. Table 3 reports the results.

Setup. Following Arjovsky et al. [4], we construct a binary classification task on MNIST digits (0–4 vs. 5–9) where color is spuriously correlated with the label. Specifically, for each digit, we assign color (red or green) based on its label with probability ρ_e in environment e. Training uses two environments with $\rho_1 = 0.9$ and $\rho_2 = 0.8$.

Table 3: Colored MNIST OOD Acc.

Method	OOD Acc (%)	ID. Stru.
ERM	27.2 $\pm$11.3	—
IRM	21.3 $\pm$13.2	$\mathcal{G}_2$ (fixed)
VREx	24.3 $\pm$12.8	—
CIRCE	10.3 $\pm$0.5	$\mathcal{G}_1$ (fixed)
SaCRL	**45.2** $\pm$10.0	$\mathcal{G}_1$ (92%)

7.4 Experiment 3: Structure Analysis on PACS

Table 4: Structure selected on PACS; each row trains on the other three domains (5 seeds). Margin: $\alpha_{(1)} - \alpha_{(2)}$; p: permutation p-value of $\min_k V_k$; Seeds: agreement of $\hat{k}$ across seeds; Cal. acc.: identification accuracy of synthetic runs in the same margin bin (Appendix H.6).

Held-out	α_1	α_2	α_3	$\hat{k}$	Margin	p	Seeds	Cal. acc. (%)
Photo	0.21	**0.68**	0.11	$\mathcal{G}_2$	0.47	0.008	5/5	86.1
Art Painting	**0.59**	0.31	0.10	$\mathcal{G}_1$	0.28	0.021	5/5	65.0
Cartoon	0.35	**0.52**	0.13	$\mathcal{G}_2$	0.17	0.037	4/5	65.0
Sketch	**0.63**	0.25	0.12	$\mathcal{G}_1$	0.38	0.014	5/5	86.1

Table 4 reports the adaptive weights and the confidence measures of Section 5 for each leave-one-domain-out split of PACS [22] (dataset details in Appendix H.3). Since PACS has no ground-truth DAG, we make three verifiable claims. (i) *The selection responds to the data*: two different structures are selected across splits, so SaCRL does not return a default. (ii) *Confidence tracks the margin*: seeds agree 5/5 on the three splits with margin at least 0.28 and 4/5 on Cartoon, the split with the smallest margin; on synthetic data with known structure, the margins of Photo and Sketch fall in a bin with 86.1% identification accuracy, and those of Art and Cartoon in a bin with 65.0%. (iii) *No canonical invariance holds exactly*: $p < 0.05$ on every split, so $\hat{k}$ is the closest canonical structure for each training set, and the calibrated accuracies in (ii), which assume that the true structure is a candidate, do not apply directly.

7.5 Experiment 4: Domain Generalization on DomainBed

Setup. We further evaluate SaCRL on DomainBed [16], a benchmark suite to test how well models generalize to unseen domains. Besides PACS [22] (4 domains, 7 classes), we choose two additional benchmarks: VLCS [45] (4 domains: Caltech101, LabelMe, SUN09, VOC2007; 5 classes), and OfficeHome [48] (4 domains: Art, Clipart, Product, Real-World; 65 classes). Each spans diverse domain shifts in style, viewpoint, and resolution. We follow the leave-one-domain-out protocol with the model of Section 7.1. Freezing the backbone holds representation capacity fixed, so differences across methods trace to the objective rather than to encoder optimization.

Table 5: Results on DomainBed benchmarks: average leave-one-domain-out OOD accuracy (%).

	PACS					**VLCS**					**OfficeHome**				
Method	Photo	Art	Cartoon	Sketch	Avg	Caltech	LabelMe	SUN09	VOC	Avg	Art	Clipart	Product	Real	Avg
ERM	93.2	66.0	51.8	37.6	62.2	94.0	60.2	69.9	73.0	74.3	49.9	41.7	66.7	67.7	56.5
IRM	93.2	66.2	51.7	37.6	62.2	95.0	60.1	70.1	72.6	74.5	49.0	41.2	66.5	67.7	56.1
VREx	93.0	65.5	51.6	37.9	62.0	93.3	60.2	69.8	71.0	73.6	48.6	41.2	66.2	67.6	55.9
CIRCE	92.5	64.5	51.4	37.3	61.4	93.8	60.0	67.7	69.7	72.8	49.0	41.9	66.0	67.4	56.1
SaCRL	**94.0**	**68.5**	**54.0**	**40.5**	**64.3**	**97.7**	**64.4**	**75.7**	**77.1**	**78.7**	**52.9**	**46.5**	**69.1**	**69.8**	**59.6**

Table 5 shows the comparison results with the baselines. SaCRL achieves the best per-domain and average accuracy on all benchmarks, with average improvements of +2.1pp (PACS), +4.2pp (VLCS), and +3.1pp (OfficeHome) over the best baseline. Fixed-invariance baselines (IRM, VREx, CIRCE) show no consistent improvement over ERM, consistent with the selected structure varying across splits (Table 4). With an end-to-end fine-tuned ResNet18 (PACS, Photo held out), all methods improve by 1.0–2.0pp and SaCRL is on par with ERM and IRM (Appendix H.13).

7.6 Experiment 5: Robustness to Misspecification and Limited Diversity

We stress-test SaCRL when (a) the true DGP falls outside the three canonical structures and (b) environment diversity is reduced. We use a bidirectional feedback DGP $X \leftrightarrow Y$ with mixing coefficient $\nu \in [0, 1]$: ν=0 is causal $X \to Y$ ($\mathcal{G}_2$), ν=1 is anti-causal $Y \to X$ ($\mathcal{G}_1$), and intermediate values fit none of $\mathcal{G}_1$–$\mathcal{G}_3$. Diversity is controlled by $\delta_{\min}$ (Definition C.18).

SaCRL outperforms every evaluated baseline in each setting (Table 6). Near the endpoints it selects the nearer structure ($\mathcal{G}_2$ at ν=0.25, $\mathcal{G}_1$ at ν=0.75). At ν=0.50, where no candidate holds, the weights stay spread ($\alpha_1 \approx \alpha_2 \approx 0.42$, $\alpha_3 \approx 0.16$): SaCRL trains with a soft model average instead of committing to a wrong structure, and still exceeds ERM by 12.7pp. A finer sweep (Appendix H.5) shows α_1 increasing and α_2 decreasing monotonically in ν, crossing near ν=0.50, with SoftMin

Table 6: Robustness under misspecification (ν) and limited diversity ($\delta_{\min}$). OOD accuracy (%) over 5 seeds. ID. Stru.: majority structure (seeds agreeing).

Setting	ERM	IRM	VREx	CIRCE	SaCRL	ID. Stru.
Misspecification ($\delta_{\min}$=1.0):						
ν=0.25 (near-causal)	34.4 ±2.7	—	—	—	**50.2** ±3.7	$\mathcal{G}_2$ (4/5)
ν=0.50 (bidirectional)	61.2 ±3.8	61.3 ±4.2	61.4 ±4.4	66.5 ±2.6	**73.9** ±4.1	$\mathcal{G}_2$ (5/5)
ν=0.75 (near-anticausal)	98.5 ±0.6	—	—	—	**99.6** ±0.3	$\mathcal{G}_1$ (4/5)
Limited diversity (ν=0.50):						
$\delta_{\min}$=1.0 (full)	61.2 ±3.8	61.3 ±4.2	61.4 ±4.4	66.5 ±2.6	**73.9** ±4.1	$\mathcal{G}_2$ (5/5)
$\delta_{\min}$=0.5 (moderate)	56.9 ±2.5	—	—	—	**67.2** ±1.8	$\mathcal{G}_2$ (5/5)
$\delta_{\min}$=0.2 (low)	53.1 ±1.7	—	—	—	**59.7** ±3.7	$\mathcal{G}_2$ (4/5)
$\delta_{\min}$=0.05 (minimal)	50.2 ±2.0	—	—	—	**57.2** ±3.7	$\mathcal{G}_2$ (5/5)

outperforming hard $\arg\min$ selection throughout and by the widest margin at the crossover. Under reduced diversity, the selected structure remains stable, and the gain over ERM persists at $\delta_{\min}$=0.05, consistent with Assumption 1 governing the informativeness of the penalty rather than its feasibility.

7.7 Ablation Studies

We conduct ablations on synthetic $\mathcal{G}_2$ data with $n = 2000$ samples to isolate the contribution of each component. Results are shown in Table 7.

Table 7: Ablation on $\mathcal{G}_2$ (n=2000). Default: SoftMin, β annealing 1→50, λ=1.0, ρ=0.5.

Variant	Stru. ID (%)	OOD Acc (%)
Full method	89.0	84.2
Hard $\arg\min_k$	79.5	79.8
Fixed $\beta = 10$	83.0	81.5
Fixed $\lambda = 0.1$	72.5	78.3
Fixed $\lambda = 5.0$	88.5	81.0
Uniform (ρ=1)	88.0	82.1
Worst-case (ρ=1/$\lvert\mathcal{E}\rvert$)	87.5	81.4

Replacing SoftMin with hard $\arg\min_k$ selection during training lowers identification accuracy from 89.0% to 79.5%: early in training, violation estimates are noisy, and hard selection commits prematurely to whichever structure has the smallest estimate. SoftMin delays commitment until the representation is informative; the reported structure is still the hard decision $\hat{k}$ made after training.

The effect of varying the number of training environments m is reported in Appendix H.11, where we observe that $m \geq 3$ yields reliable structure identification while $m = 2$ is unstable. Hyperparameter sensitivity sweeps and a cross-fitting variant of conditional HSIC are reported in Appendix H.4 and Appendix H.8, respectively. Computational efficiency analysis (training time, memory, runtime breakdown) is in Appendix H.12.

8 Conclusion

We proposed a structure-agnostic approach to causal representation learning that formulates causal structure selection as a soft optimization problem over HSIC-based invariance violation metrics. The key insight is that, when the invariant component of X is predictive, correct invariances can be achieved with zero violation while preserving predictive accuracy, whereas incorrect invariances cannot, enabling automatic structure identification through optimization. We developed (i) a unified measure-theoretic framework that extends prior work on anti-causal learning to confounded-descendant and confounded-outcome settings; (ii) an efficient learning algorithm whose complexity scales linearly with the number of candidate structures, leveraging SoftMin aggregation and RFF-based HSIC approximations; and (iii) theoretical guarantees for structure identification, invariance satisfaction, and OOD generalization. Empirically, SaCRL recovers the true structure in most runs on synthetic and semi-synthetic Bayesian-network benchmarks, with a weight margin that tracks identification accuracy, surpasses fixed-invariance baselines on Colored MNIST and three DomainBed benchmarks, and outperforms ERM under bidirectional dynamics.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their constructive comments. This work was supported in part by the National Science Foundation under Grant Nos. ECCS-2216926, CCF-2331302, CNS-2241713, and CNS-2339686.

References

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 2021.

[2] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, 2023.

[3] Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *Artificial Intelligence and Statistics*, 2024.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[5] Arman Behnam and Binghui Wang. Graph neural network causal explanation via neural causal models. In *European conference on computer vision*, pages 410–427. Springer, 2024.

[6] Arman Behnam and Binghui Wang. Measure-theoretic anti-causal representation learning. *Advances in Neural Information Processing Systems*, 38:61375–61431, 2025.

[7] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 2022.

[8] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

[10] Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems*, 2022.

[11] Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. doi: 10.1109/ICCV.2013.208.

[13] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, volume 3734 of *Lecture Notes in Computer Science*, pages 63–77. Springer, 2005.

[14] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, pages 585–592, 2007.

[15] Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. Joint learning of label and environment causal independence for graph out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[17] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

[18] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[19] Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain shifts. *Advances in Neural Information Processing Systems*, 35:20782–20794, 2022.

[20] Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. *arXiv preprint arXiv:2206.07837*, 2022.

[21] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 2021.

[22] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *IEEE international conference on computer vision*, 2017.

[23] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 2022.

[24] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34:6155–6170, 2021.

[25] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[26] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

[27] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[28] Maggie Makar and Alexander D'Amour. Fairness and robustness in anti-causal prediction. *arXiv preprint arXiv:2209.09423*, 2022.

[29] Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. *Advances in neural information processing systems*, 31, 2018.

[30] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.

[31] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

[32] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[33] Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with xrm. *arXiv preprint arXiv:2309.16748*, 2023.

[34] Roman Pogodin, Namrata Deka, Yazhe Li, Danica J Sutherland, Victor Veitch, and Arthur Gretton. Efficient conditionally invariant representation learning. *arXiv*, 2022.

[35] Roman Pogodin, Antonin Schrab, Yazhe Li, Danica J Sutherland, and Arthur Gretton. Practical kernel tests of conditional independence. *arXiv preprint arXiv:2402.13196*, 2024.

[36] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[37] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 28656–28679. PMLR, 2023.

[38] Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. Learning neural causal models with active interventions. *arXiv preprint arXiv:2109.02429*, 2021.

[39] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[40] Jia Shi, Gautam Gare, Jinjin Tian, Siqi Chai, Zhiqiu Lin, Arun Vasudevan, Di Feng, Francesco Ferroni, and Shu Kong. Lca-on-the-line: Benchmarking out-of-distribution generalization with class taxonomies. *arXiv preprint arXiv:2407.16067*, 2024.

[41] Chandler Squires and Caroline Uhler. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 23(5):1781–1815, 2023.

[42] Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, 2023.

[43] Yongduo Sui, Qitian Wu, Jiancan Wu, Qing Cui, Longfei Li, Jun Zhou, Xiang Wang, and Xiangnan He. Unleashing the power of graph data augmentation on covariate distribution shift. *Advances in Neural Information Processing Systems*, 36, 2024.

[44] Zoltán Szabó and Bharath K Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

[45] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[46] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[47] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

[48] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[49] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, 2023.

[50] Xin Wang, Hong Chen, Zihao Wu, Wenwu Zhu, et al. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[51] Zihao Wang and Victor Veitch. The causal structure of domain invariant supervised representation learning. *arXiv preprint arXiv:2208.06987*, 2023.

[52] Liang Wendong, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. *Advances in Neural Information Processing Systems*, 36:32481–32520, 2023.

[53] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[54] Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems*, 35:12964–12978, 2022.

[55] Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems*, 2023.

[56] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5372–5382, 2021.

[57] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13025–13032, 2020.

[58] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, pages 561–578. Springer, 2020.

Appendix

A Related Work **17**

B Notation **18**

C Mathematical Preliminaries **18**

C.1 Measure-Theoretic Foundations 18

C.2 Causal Structures and Directed Acyclic Graphs 19

C.3 Interventions and the Do-Calculus 19

C.4 Measure-Theoretic Causality 20

C.5 Kernel Independence Framework Definitions 21

C.6 Assumptions 21

D Theoretical Results **22**

D.1 Violation Characterization, Structure Identification & Minimizers of the Objective . 22

D.2 Generalization Bounds 24

E Proofs **25**

E.1 Proof of Lemma 1 (HSIC Characterizes Independence) 26

E.2 Proof of Lemma 2 (Conditional HSIC Consistency) 26

E.3 Proof of Lemma 3 (SoftMin Properties) 27

E.4 Proof of Theorem D.1 (Violation Characterization) 28

E.5 Proof of Theorem D.2 28

E.6 Proof of Corollary 1 28

E.7 Proof of Corollary 2 29

E.8 Proof of Theorem D.3 29

E.9 Proof of Proposition 1 (Minimizers of the Objective) 29

E.10 Proof of Theorem D.4 (Generalization Bound) 30

E.11 Proof of Theorem D.5 (Sample Complexity) 30

E.12 Proof of Proposition 2 (Differentiable CVaR) 31

E.13 Proof of Theorem F.1 (Algorithm Convergence) 31

E.14 Proof of Theorem G.1 (Heterogeneous Structure Identification) 32

F Algorithm Details **32**

F.1 Main Algorithm 32

F.2 Subroutines 33

F.3 Complexity Analysis 34

F.4 Convergence Guarantees 34

G Extension to Heterogeneous Structures **35**

H Additional Experiments **36**
H.1 Implementation 36
H.2 Synthetic DGP Equations 36
H.3 Real-World Datasets 37
H.4 Hyperparameter Sensitivity Results 37
H.5 Crossover Between Candidate Structures 37
H.6 Margin Calibration 37
H.7 Semi-Synthetic Bayesian-Network Benchmarks 37
H.8 Cross-Fitting Variant 38
H.9 EMA Conditioning Target 38
H.10 RFF Approximation Error 39
H.11 Effect of the Number of Training Environments 39
H.12 Computational Efficiency and Runtime Analysis 39
H.13 Fine-Tuned Backbone 39

I Limitation **40**

A Related Work

Distribution-Invariant Learning. A large body of work in domain adaptation and domain generalization aims to learn representations that are invariant across environments or distributions. Early non-causal approaches such as CDAN [25] align feature distributions across domains, while DDAIG [57] synthesizes new domains by adversarial image generation. More recent methods, including FeAT [11] and AIA [43], design more expressive objectives to improve robustness under distribution shift.

Representative domain generalization methods, evaluated on benchmarks such as DomainBed [16], include ERM [46], L2A-OT [58], StableNet [56], SagNets [30], SWAD [9], FACT [53], Ratatouille [37], XRM [33], and IRM [4], along with extensions such as REx [21] and information bottleneck approaches [1]. While effective in practice, these methods generally enforce distributional invariance without explicitly reasoning about the underlying causal structure or validating whether the assumed invariance is causally appropriate. Counterfactual invariance [47] and transportable representations [19] do reason about the causal structure, but they take it as given. Weight-averaging methods such as SWAD [9] modify the optimization trajectory rather than the objective, and are complementary to SaCRL, which determines which invariance is enforced.

Intervention-based Causal Learning. Intervention-based causal representation learning leverages environment changes or interventions to identify causal factors. Foundational works include nonparametric ICR [49], which jointly learns representations and intervention targets under SCM assumptions, and learning under general nonlinear mixing [8], which identifies linear latent causal models from interventions.

Several weakly supervised approaches [7, 26, 27, 40] exploit partial intervention information to learn invariant representations, while ICRL [2] extends these ideas to imperfect interventions. Weak distributional invariance [3] studies identifiability under perfect and imperfect interventions by identifying latent variables with stable distributional properties.

Unsupervised approaches such as ICA-based causal learning [52] aim to recover latent causal variables through component analysis and statistical independence criteria. AIT [38] learns the causal graph of an explicit SCM by actively selecting interventions. These methods typically rely on either explicit intervention signals or strong assumptions about the form of the causal graph.

Structure-based Causal Representation Learning. Structure-based causal representation learning methods explicitly assume or model a particular causal structure. Wang and Veitch [51] characterize the anti-causal, confounded-descendant, and confounded-outcome structures on which SaCRL builds and derive the invariance appropriate for each, and CACM [20] adapts the invariance constraint to a given causal graph; both require the structure as input, which SaCRL instead selects from data. Causal Disentanglement [42] learns representations aligned with latent causal graphs, while classical approaches such as ICP [31] and its nonlinear extension [17] exploit invariance across environments to identify causal mechanisms.

Related works such as CSG [24], which separates semantic from variation factors with a causal generative model, and LECI [15], which identifies causal subgraphs in graph data, remove spurious correlations under DAG assumptions. KCDC [29] orients cause–effect pairs by comparing the variability of kernel conditional mean embeddings. Anti-causal modeling approaches [28] study settings where inputs are generated from labels and protected attributes using conventional causal models and do-calculus, while Behnam and Wang [6] develop measure-theoretic formulations for anti-causal representation learning.

Research Gap. Existing approaches face a trilemma. Structure-based methods achieve strong guarantees but require knowing the causal graph. Invariant learning methods avoid explicit graphs but hardcode implicit structural assumptions that cause failures under mismatch. Causal discovery methods learn structure but decouple it from representation learning, preventing joint optimization. No existing framework selects among candidate causal structures and learns a representation adapted to the selected structure in a unified, end-to-end differentiable objective. Our work fills this gap: we introduce a structure-agnostic objective that evaluates representations against all candidate causal structures, automatically concentrating on the correct one through adaptive weighting, while jointly optimizing predictive performance.

B Notation

Table 8 summarizes notation used throughout. We organize symbols into six categories: spaces and sets defining the mathematical objects; random variables and observed data; causal graph components encoding structural assumptions; kernels and independence testing machinery; learning components including representations, predictors, and losses; and theoretical quantities appearing in our analysis. Lowercase denotes realizations (x, y), uppercase denotes random variables (X, Y), and calligraphic denotes spaces ($\mathcal{X}, \mathcal{Y}$). Subscripts index environments (P_e), structures ($\mathcal{G}_k$), components ($X_z, X_z^\perp$), and iterates (ϕ_t).

Table 8: Notation summary.

Spaces & Sets		**Random Variables & Data**	
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	Input, output, representation spaces	X, Y, E	Features, target, environment
$\mathcal{E}$	Environment set $\{e_1, \ldots, e_m\}$	U, Z	Latent confounder, mediator
$(\Omega, \mathscr{F}, P)$	Probability space	$X_z, X_z^\perp$	Mediated, orthogonal components
$\mathcal{H}, \mathcal{H}_\mathcal{X}$	RKHS, RKHS on $\mathcal{X}$	D_e	Dataset $\{(x_j^e, y_j^e)\}_{j=1}^{n_e}$
Δ^2	Probability simplex	n, m, d	Samples, environments, dimension
Causal Structure		**Kernels & Independence**	
$\mathcal{G} = (V, E)$	DAG with vertices, edges	$k, k_\mathcal{X}$	Reproducing kernel
$\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$	Anti-causal, conf-desc, conf-out	κ	Kernel bound: $k(x,x) \leq \kappa^2$
$\mathcal{G}^* = \mathcal{G}_{k^*}$	True structure	μ_P	Mean embedding of P
$\mathrm{Pa}(v)$	Parents of v	$\mathcal{C}_{XY}$	Cross-covariance operator
$K_S, K_S^\mathbb{Q}$	Causal, interventional kernel	$\mathrm{HSIC}(X, Y)$	Independence criterion
$\mathcal{I}_k$	Invariance for structure k	$\tilde{X}, \tilde{Y}$	Regression residuals
Learning Components		**Theoretical Quantities**	
ϕ, w	Representation, predictor	$\mathbf{D}$	Diversity matrix ($m \times 3$)
$\Phi, \mathcal{W}$	Function classes	$\delta_{\min}$	Min diversity parameter
$\ell, \mathcal{R}_e, \hat{\mathcal{R}}_e$	Loss, population/empirical risk	η	Faithfulness gap
V_k, V_k^*	Empirical/population violation	$d_\mathcal{K}(e, e')$	Interventional distance
α_k	Adaptive weight $\propto e^{-\beta V_k}$	ξ_n	Estimation error
$\lambda, \beta, \rho, \gamma$	Regularization, temp, CVaR, ridge	L_w	Lipschitz constant
$\mathrm{Ent}(\boldsymbol{\alpha})$	Entropy of the weights	ϵ_0	Predictiveness tolerance

C Mathematical Preliminaries

C.1 Measure-Theoretic Foundations

Definition C.1 (Measurable Space)**.** A pair $(\Omega, \mathscr{F})$ is a measurable space where Ω is a non-empty set and $\mathscr{F}$ is a σ-algebra on Ω, i.e., a collection of subsets satisfying: (i) $\emptyset, \Omega \in \mathscr{F}$; (ii) $A \in \mathscr{F} \Rightarrow A^c \in \mathscr{F}$; (iii) $\{A_i\}_{i=1}^\infty \subset \mathscr{F} \Rightarrow \bigcup_{i=1}^\infty A_i \in \mathscr{F}$.

Definition C.2 (Probability Measure)**.** A probability measure on a measurable space $(\Omega, \mathscr{F})$ is a function $\mu : \mathscr{F} \to [0, 1]$ satisfying $\mu(\Omega) = 1$. For any countable collection $\{A_i\}_{i=1}^\infty$ of pairwise disjoint sets in $\mathscr{F}$,

$$\mu\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mu(A_i) \tag{14}$$

The triple $(\Omega, \mathscr{F}, \mu)$ is a probability space.

Definition C.3 (Product Space)**.** Given measurable spaces $(\Omega_1, \mathscr{F}_1)$ and $(\Omega_2, \mathscr{F}_2)$, the product σ-algebra $\mathscr{F}_1 \otimes \mathscr{F}_2$ on $\Omega_1 \times \Omega_2$ is the σ-algebra generated by measurable rectangles $A_1 \times A_2$ where $A_1 \in \mathscr{F}_1$ and $A_2 \in \mathscr{F}_2$. For probability measures μ_1 on $(\Omega_1, \mathscr{F}_1)$ and μ_2 on $(\Omega_2, \mathscr{F}_2)$, the product measure $\mu_1 \otimes \mu_2$ on $(\Omega_1 \times \Omega_2, \mathscr{F}_1 \otimes \mathscr{F}_2)$ is uniquely determined by $(\mu_1 \otimes \mu_2)(A_1 \times A_2) = \mu_1(A_1) \cdot \mu_2(A_2)$ for all $A_1 \in \mathscr{F}_1$, $A_2 \in \mathscr{F}_2$.

Definition C.4 (Markov Kernel)**.** A (Markov) kernel from a measurable space $(\Omega_1, \mathscr{F}_1)$ to a measurable space $(\Omega_2, \mathscr{F}_2)$ is a mapping $K : \Omega_1 \times \mathscr{F}_2 \to [0, 1]$. For each fixed $\omega \in \Omega_1$, the function $A \mapsto K(\omega, A)$ is a probability measure on $(\Omega_2, \mathscr{F}_2)$. For each fixed $A \in \mathscr{F}_2$, the function $\omega \mapsto K(\omega, A)$ is $\mathscr{F}_1$-measurable.

Intuitively, $K(\omega, A)$ represents the conditional probability of event $A \in \mathscr{F}_2$ given state $\omega \in \Omega_1$.

C.2 Causal Structures and Directed Acyclic Graphs

Definition C.5 (Directed Acyclic Graph)**.** A directed acyclic graph (DAG) is a pair $\mathcal{G} = (V, E)$ where V is a finite set of vertices (nodes) and $E \subseteq V \times V$ is a set of directed edges such that there exists no sequence of vertices $v_1, \ldots, v_k$ with $k \geq 2$, $v_1 = v_k$, and $(v_i, v_{i+1}) \in E$ for all $i = 1, \ldots, k-1$.

Definition C.6 (Causal Structure)**.** A causal structure is a tuple $\mathcal{S} = (V, E, \{\mathscr{F}_v\}_{v \in V}, \{K_v\}_{v \in V})$ where (V, E) is a DAG with vertices representing random variables. For each $v \in V$, $(\Omega_v, \mathscr{F}_v)$ is a measurable space. For each $v \in V$, K_v is a kernel from $\left(\prod_{u \in \mathrm{Pa}(v)} \Omega_u, \bigotimes_{u \in \mathrm{Pa}(v)} \mathscr{F}_u\right)$ to $(\Omega_v, \mathscr{F}_v)$ representing the causal mechanism by which parents generate the variable v.

For root nodes with $\mathrm{Pa}(v) = \emptyset$, the kernel K_v reduces to a probability measure. We now formalize how observed features arise from multiple generating mechanisms.

Definition C.7 (Causal Feature Decomposition)**.** Let Z be a latent variable in the causal graph $\mathcal{G}$. The observed features X admit a causal decomposition with respect to Z if there exist measurable functions $\pi_z : \mathcal{X} \to \mathcal{X}_z$ and $\pi_z^\perp : \mathcal{X} \to \mathcal{X}_z^\perp$ such that $X = (\pi_z(X), \pi_z^\perp(X)) =: (X_z, X_z^\perp)$. X_z contains all and only the information in X that is causally downstream of Z, and $X_z^\perp$ contains the remaining information in X that is not mediated by Z.

We write $X = (X_z, X_z^\perp)$ to denote this decomposition, where $\mathcal{X} \cong \mathcal{X}_z \times \mathcal{X}_z^\perp$.

C.3 Interventions and the Do-Calculus

The distinction between observational conditioning $P(Y \mid X = x)$ and interventional conditioning $P(Y \mid \mathrm{do}(X = x))$, where the latter severs incoming edges to X in the causal graph, explains why different causal structures call for different invariances. Our method itself does not require do-calculus.

Interventions distinguish causal relationships from statistical associations. We formalize both perfect (hard) and imperfect (soft) interventions within the generalized framework.

Definition C.8 (Perfect and Imperfect Interventions)**.** An intervention on a causal structure $\mathcal{S}$ is specified by a measurable mapping $\mathbb{Q} : \Omega \times \mathscr{F} \to [0, 1]$ where $\mathscr{F}$ is a suitable σ-algebra.

(i) A perfect (hard) intervention on variable V replaces its causal mechanism K_V with a fixed distribution $\mathbb{Q}$ independent of parent values: $\mathbb{Q}(A \mid \omega) = \mathbb{Q}(A)$ for all $\omega \in \Omega$, $A \in \mathscr{F}$

(ii) An imperfect (soft) intervention on variable V modifies its causal mechanism to depend on both parents and intervention parameters: $\mathbb{Q}(A \mid \omega)$ varies with $\omega \in \Omega$

The do-operator $\mathrm{do}(X = x)$ corresponds to a perfect intervention that sets X to value x by replacing the kernel K_X with a point mass at x, formally: $\mathbb{Q}(A) = \mathbb{I}_{\{x\}}(A)$ for measurable sets A.

Definition C.9 (Causal Independence)**.** Random variables X and Y are causally independent given Z, denoted $X \perp\!\!\!\perp_c Y \mid Z$, if

$$P(Y \mid \mathrm{do}(X = x), Z = z) = P(Y \mid Z = z) \tag{15}$$

for all x in the support of X and all z in the support of Z, where $\mathrm{do}(X = x)$ denotes an intervention setting X to value x.

The do-operator $\mathrm{do}(X = x)$ differs fundamentally from $P(Y \mid X = x)$ which observes $X = x$ while preserving causal relationships, whereas $P(Y \mid \mathrm{do}(X = x))$ forces $X = x$ by intervention.

C.4 Measure-Theoretic Causality

The causal-space, product-causal-space, and interventional-kernel constructions below follow Behnam and Wang [6]; we extend them with structure-indexed kernel families (Definition C.14).

Definition C.10 (Sub-σ-algebra)**.** Given a product space $\Omega = \prod_{t\in T} \Omega_t$ with index set T and σ-algebras $\mathscr{F}_t$ on Ω_t, for any subset $S \subseteq T$, the sub-σ-algebra $\mathscr{F}_S$ is the σ-algebra generated by measurable rectangles of the form $\prod_{t\in S} A_t \times \prod_{t\in T\setminus S} \Omega_t$ where $A_t \in \mathscr{F}_t$ for $t \in S$.

Intuitively, $\mathscr{F}_S$ captures events that depend only on the components indexed by S.

Definition C.11 (Causal Kernel)**.** A causal kernel for index set $S \subseteq T$ is a function $K_S : \Omega \times \mathscr{F} \to [0, 1]$. For each fixed $\omega \in \Omega$, the mapping $A \mapsto K_S(\omega, A)$ is a probability measure on $(\Omega, \mathscr{F})$. For each fixed $A \in \mathscr{F}$, the mapping $\omega \mapsto K_S(\omega, A)$ is $\mathscr{F}_S$-measurable.

Definition C.12 (Causal Space)**.** For an environment $e \in \mathcal{E}$, a causal space is a tuple $(\Omega_e, \mathscr{F}_e, P_e, \mathbb{K}_e, \mathcal{G}_e)$ where $\Omega_e = \prod_{v\in V_e} \Omega_v$ is the sample space with product structure, and V_e is the vertex set of the causal DAG and each Ω_v is the state space of variable v. Here, $\mathscr{F}_e = \bigotimes_{v\in V_e} \mathscr{F}_v$ is the product σ-algebra, where $\mathscr{F}_v$ is the σ-algebra on Ω_v, P_e is a probability measure on $(\Omega_e, \mathscr{F}_e)$ compatible with $\mathcal{G}_e$, $\mathbb{K}_e = \{K_v^{(e)}\}_{v\in V_e}$ is the family of causal mechanisms. Each $K_v^{(e)}$ is a Markov kernel (Definition C.4) from the parent space $\left(\prod_{u\in \mathrm{Pa}(v)} \Omega_u, \bigotimes_{u\in \mathrm{Pa}(v)} \mathscr{F}_u\right)$ to $(\Omega_v, \mathscr{F}_v)$, encoding the conditional distribution $P_e(V = \cdot \mid \mathrm{Pa}(v) = \cdot)$, and $\mathcal{G}_e = (V_e, E_e)$ is the causal DAG specifying the qualitative causal structure.

The joint distribution factorizes according to the DAG:

$$P_e(\omega) = \prod_{v\in V_e} K_v^{(e)}\left(\omega_{\mathrm{Pa}(v)}, \{\omega_v\}\right) \tag{16}$$

where $\omega = (\omega_v)_{v\in V_e} \in \Omega_e$ and $\omega_{\mathrm{Pa}(v)}$ denotes the components corresponding to parents of v.

Definition C.13 (Product Causal Space)**.** Given causal spaces $\{(\Omega_{e_i}, \mathscr{F}_{e_i}, P_{e_i}, K_{e_i}, \mathcal{G}_{e_i})\}_{i=1}^m$ for environments $\{e_i\}_{i=1}^m$, the product causal space is the tuple $(\Omega, \mathscr{F}, \mathbb{P}, \mathbb{K}, \mathcal{G})$ where $\Omega = \prod_{i=1}^m \Omega_{e_i}$ is the product sample space, $\mathscr{F} = \bigotimes_{i=1}^m \mathscr{F}_{e_i}$ is the product σ-algebra, $\mathbb{P} = \bigotimes_{i=1}^m P_{e_i}$ is the product measure, $\mathbb{K} = \{K_S : S \in \mathscr{P}(T)\}$ is a family of causal kernels indexed by subsets S of the combined index set $T = \bigcup_{i=1}^m V_{e_i}$ of all variables, and $\mathcal{G} = (V, E)$ is a causal DAG compatible with the individual environment graphs.

The product causal space admits sub-σ-algebras and causal kernels indexed by variable subsets; see Definitions C.10 and C.11.s.

Definition C.14 (Structure-Agnostic Causal Space)**.** A structure-agnostic causal space extends the product causal space (Definition C.13) with structure-indexed components: $(\Omega, \mathscr{F}, \mathbb{P}, \{\mathbb{K}_k\}_{k=1}^3, \{\mathcal{G}_k\}_{k=1}^3)$ where $(\Omega, \mathscr{F}, \mathbb{P})$ is the product space over environments, $\{\mathcal{G}_k\}_{k=1}^3$ are the three candidate causal DAGs (Definition 2.1), and $\mathbb{K}_k = \{K_S^{(k)} : S \subseteq T\}$ is the causal kernel family encoding how interventions propagate under structure $\mathcal{G}_k$.

The $\mathscr{F}_S$-measurability condition ensures $K_S(\omega, A)$ depends only on components indexed by S, encoding graph-specified causal dependencies.

Definition C.15 (Interventional Kernel)**.** Let $(\Omega, \mathscr{F}, \mathbb{P}, \mathbb{K}, \mathcal{G})$ be a product causal space with causal kernel K_S for index set $S \subseteq T$. Let $\mathbb{Q} : \Omega \times \mathscr{F} \to [0, 1]$ be an intervention kernel, where $\mathbb{Q}(\omega', A)$ specifies the probability of event A when intervening at state ω'.

The interventional kernel $K_S^{\mathbb{Q}} : \Omega \times \mathscr{F} \to [0, 1]$ is defined by:

$$K_S^{\mathbb{Q}}(\omega, A) := \int_\Omega \mathbb{Q}(\omega', A)\, K_S(\omega, d\omega') \tag{17}$$

where $\omega \in \Omega$ is the conditioning state (values of variables in S), $K_S(\omega, d\omega')$ is the original causal kernel giving the distribution over ω' given ω, and $\mathbb{Q}(\omega', A)$ is the intervention's effect on event A at state ω'.

This construction unifies perfect and imperfect interventions: when $\mathbb{Q}(A \mid \omega') = \mathbb{Q}(A)$ is constant, the interventional kernel corresponds to a hard intervention; when $\mathbb{Q}(A \mid \omega')$ varies with ω', it represents a soft intervention.

C.5 Kernel Independence Framework Definitions

Definition C.16 (Reproducing Kernel Hilbert Space)**.** Let $\mathcal{X}$ be a non-empty set. A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is a reproducing kernel Hilbert space (RKHS) if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that:

1. $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$
2. $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, $x \in \mathcal{X}$ (reproducing property)

Definition C.17 (Cross-Covariance Operator)**.** Let $(\mathcal{X}, \mathcal{H}_\mathcal{X}, k_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{H}_\mathcal{Y}, k_\mathcal{Y})$ be RKHS pairs. The cross-covariance operator $\mathcal{C}_{XY} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{X}$ is defined by

$$\mathcal{C}_{XY} := \mathbb{E}_{XY}[(k_\mathcal{X}(\cdot, X) - \mu_X) \otimes (k_\mathcal{Y}(\cdot, Y) - \mu_Y)] \tag{18}$$

where $\otimes$ denotes the tensor product.

C.6 Assumptions

Here we state assumptions required for theoretical findings.

Assumption 2 (Characteristic Kernels)**.** The kernels $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, $k_\mathcal{Z} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, and $k_\mathcal{E} : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ are characteristic and bounded: $\sup_x k(x, x) \le \kappa^2 < \infty$.

Characteristic kernels ensure HSIC fully characterizes independence (Lemma 1). We further require the tensor products of these kernels used in HSIC to be characteristic, which holds for Gaussian and delta kernels [44]. For the conditional HSIC tests underlying V_1 and V_2, we additionally require the regression operator to be well-behaved.

Assumption 3 (Kernel Regularity)**.** The kernel regression underlying V_1 and V_2 satisfies:

1. The regression functions $z \mapsto \mathbb{E}[A \mid Z = z]$ estimated in V_1 and V_2 lie in the RKHS of the kernel on the conditioning variable Z.
2. The eigenvalues $\{\lambda_j\}$ of the kernel integral operator satisfy $\sum_j \lambda_j^{1/2} < \infty$.
3. The regressed variables follow the additive-noise model of Lemma 2.

Beyond kernel conditions, structure identification requires environments that vary in distinguishable ways. We formalize this via a diversity matrix measuring distributional shifts across environments.

Definition C.18 (Environment Diversity Matrix)**.** Let $\mathcal{E} = \{e_1, \ldots, e_m\}$ be the training environments and $\bar{P}$ the pooled distribution. Center each kernel mean embedding at its average over environments: $\bar{\mu}_X := \frac{1}{m}\sum_e \mu_{P_e(X)}$, $\bar{\mu}_{X|Y=y} := \frac{1}{m}\sum_e \mu_{P_e(X|Y=y)}$, and $\bar{\mu}_{Y|X=x} := \frac{1}{m}\sum_e \mu_{P_e(Y|X=x)}$. The $m \times 3$ diversity matrix $\mathbf{D}$ has entries

$$\begin{aligned} D_{e1} &:= \|\mu_{P_e(X)} - \bar{\mu}_X\|_\mathcal{H} && \text{(marginal shift)}, \\ D_{e2} &:= \big(\mathbb{E}_{y\sim\bar{P}(Y)}\|\mu_{P_e(X|Y=y)} - \bar{\mu}_{X|Y=y}\|_\mathcal{H}^2\big)^{1/2} && \text{(conditional shift)}, \\ D_{e3} &:= \big(\mathbb{E}_{x\sim\bar{P}(X)}\|\mu_{P_e(Y|X=x)} - \bar{\mu}_{Y|X=x}\|_\mathcal{H}^2\big)^{1/2} && \text{(outcome shift)}. \end{aligned} \tag{19}$$

For characteristic kernels, column j of $\mathbf{D}$ is zero if and only if the corresponding distribution is identical across environments ($\bar{P}$-almost everywhere for $j = 2, 3$). The matching invariance is then untestable: a zero first column makes $\mathcal{I}_3$ untestable, a zero second column $\mathcal{I}_1$, and a zero third column $\mathcal{I}_2$. Assumption 1 requires every column to be nonzero, quantified by the minimum diversity parameter

$$\delta_{\min} := \min_{j\in\{1,2,3\}} \max_{e\in\mathcal{E}} D_{ej} > 0. \tag{20}$$

Full column rank, $\operatorname{rank}(\mathbf{D}) = 3$, is sufficient for $\delta_{\min} > 0$ but not necessary.

Assumption 4 (Faithfulness)**.** The joint distribution $P(X, Y, E)$ is faithful to the true causal structure $\mathcal{G}^*$: all and only the conditional independencies implied by $\mathcal{G}^*$ hold in P.

Faithfulness ensures that violations of wrong invariances are detectable. Without it, a distribution could satisfy independencies not implied by the graph, making structures indistinguishable.

Assumption 5 (Predictive Invariant Component)**.** For some $\epsilon_0 \geq 0$, the invariant component $X_z^\perp = \pi_z^\perp(X)$ of the true structure satisfies $\pi_z^\perp \in \Phi$ and $\inf_{w \in \mathcal{W}} \max_{e \in \mathcal{E}} \mathcal{R}_e(w \circ \pi_z^\perp) \leq \mathcal{R}^* + \epsilon_0$, where $\mathcal{R}^*$ is the Bayes-optimal worst-case risk (Definition D.1).

This holds with $\epsilon_0 = 0$ when X_z carries no worst-case predictive information beyond $X_z^\perp$, and fails for small ϵ_0 when non-invariant features are predictive in the training environments, as the color is in Colored MNIST. A larger ϵ_0 enlarges the class Φ_0 of Definition D.2 and thus shrinks the faithfulness gap η.

Assumption 6 (Separation)**.** The faithfulness gap is positive: $\eta > 0$ (Definition D.2).

Faithfulness implies that $X_z^\perp$ violates every wrong invariance (proof of Theorem D.1); Assumption 6 requires the same for all representations in Φ_0, with a uniform margin.

Assumption 7 (Violation Concentration)**.** For all $t > 0$ and $k \in \{1, 2, 3\}$, the empirical violations computed from n samples per environment satisfy $\mathbb{P}\big(\sup_{\phi \in \Phi_0} |V_k(\phi) - V_k^*(\phi)| > t\big) \leq 2\exp(-nt^2/(32\kappa^8))$.

For a fixed ϕ and $k = 3$, a bound of this form follows from McDiarmid's inequality (Lemma 4). For V_1, V_2 it additionally requires the kernel-regression residuals to converge (Lemma 5), and uniformity over Φ_0 requires a covering argument, which multiplies the right-hand side by a covering number of Φ_0. We state it as an assumption, as is standard for residual-based conditional HSIC. The remaining assumptions concern the function classes used for prediction and representation.

Assumption 8 (Predictor Lipschitz Continuity)**.** The predictor class $\mathcal{W}$ consists of L_w-Lipschitz functions: for all $w \in \mathcal{W}$ and $z, z' \in \mathcal{Z}$,

$$\|w(z) - w(z')\|_2 \leq L_w \|z - z'\|_2 \tag{21}$$

with $L_w \leq L_{\max} < \infty$ uniformly over $\mathcal{W}$.

This Lipschitz condition controls how prediction errors propagate through distribution shift, appearing directly in our generalization bound (Theorem D.4).

Assumption 9 (Representation Class Regularity)**.** The representation class Φ is parameterized by $\theta \in \Theta \subseteq \mathbb{R}^p$ where:

1. Θ is compact and convex,
2. $\phi_\theta(x)$ is continuously differentiable in θ,
3. The violation functions $V_k(\phi_\theta)$ are L_V-smooth in θ.

These regularity conditions are standard for neural network function classes and ensure convergence of gradient-based optimization. Assumption 1 can be checked before training: (1) estimate $\mathbf{D}$ with empirical (conditional) mean embeddings; (2) for each column j, compare $\max_e \hat{D}_{ej}$ with its permutation null obtained by shuffling environment labels; (3) if column j is not significantly nonzero, remove from the candidate set the structure whose invariance it tests (j=1: $\mathcal{G}_3$, j=2: $\mathcal{G}_1$, j=3: $\mathcal{G}_2$); (4) if no column is significant, the environments carry no information about the structure, and ERM is the appropriate fallback.

D Theoretical Results

D.1 Violation Characterization, Structure Identification & Minimizers of the Objective

The key insight is that under the true structure, the corresponding invariance is achievable without sacrificing predictive accuracy, while wrong invariances are not. The key insight is that, when the invariant component $X_z^\perp$ is sufficiently predictive (Assumption 5), the invariance of the true structure is achievable without sacrificing predictive accuracy, while wrong invariances are not. We first introduce the definitions the theorems rely on.

Definition D.1 (ϵ-Predictive Representation)**.** A representation $\phi : \mathcal{X} \to \mathcal{Z}$ is *ϵ-predictive* if:

$$\inf_{w \in \mathcal{W}} \max_{e \in \mathcal{E}} \mathcal{R}_e(w \circ \phi) \leq \mathcal{R}^* + \epsilon \tag{22}$$

where $\mathcal{R}^* := \inf_{f: \mathcal{X} \to \mathcal{Y}} \max_e \mathcal{R}_e(f)$ is the Bayes-optimal worst-case risk.

Denote the class of ϵ-predictive representations as Φ_ϵ. We define the base predictive class $\Phi_0 := \Phi_{2\epsilon_0}$, where ϵ_0 is the tolerance of Assumption 5; the factor 2 leaves room for the minimizers of the objective (Proposition 1). One needs to measure the minimum violation incurred by any predictive representation when the wrong invariance is imposed.

Definition D.2 (Faithfulness Gap)**.** Let $\mathcal{G}^* = \mathcal{G}_{k^*}$ be the true causal structure. The faithfulness gap is:

$$\eta := \min_{k \neq k^*} \inf_{\phi \in \Phi_0} V_k^*(\phi) \tag{23}$$

Faithfulness (Assumption 4) ensures that $X_z^\perp$ violates every wrong invariance; Assumption 6 extends this to a uniform margin $\eta > 0$ over Φ_0. Assumption 1 is necessary for this margin: if the first or second column of $\mathbf{D}$ vanishes, $\mathcal{I}_3$ or $\mathcal{I}_1$ holds for every representation, and if the third vanishes, $\mathcal{I}_2$ holds for the identity map (when it belongs to Φ_0), so $\eta = 0$ whenever the matching structure is not the true one. Larger η implies easier structure identification; $\eta \to 0$ indicates near-Markov equivalence between structures.

Theorem D.1 (Violation Characterization)**.** *Under Assumptions 2–6, let $\epsilon_0 \leq \epsilon \leq 2\epsilon_0$. For the class of ϵ-predictive representations Φ_ϵ:*

1. ***Achievability:*** *If $\mathcal{G}^* = \mathcal{G}_k$, then $\inf_{\phi \in \Phi_\epsilon} V_k^*(\phi) = 0$, attained at the invariant component $X_z^\perp$.*
2. ***Non-achievability:*** *If $\mathcal{G}^* \neq \mathcal{G}_k$, then $\inf_{\phi \in \Phi_\epsilon} V_k^*(\phi) \geq \eta > 0$.*

The constraint $\phi \in \Phi_\epsilon$ prevents trivial solutions: a constant representation $\phi(x) = c$ achieves $V_1 = V_3 = 0$, but has $\mathcal{R}_e(w \circ \phi) = \mathcal{R}_{\text{trivial}} \gg \mathcal{R}^* + \epsilon$, hence $\phi \notin \Phi_\epsilon$. The joint objective $\mathcal{L}$ enforces this implicitly via the prediction loss (Proposition 1).

The asymmetry established in Theorem D.1 (achievable invariance for the true structure versus irreducible violations for wrong structures) directly enables consistent identification from finite samples.

Theorem D.2 (Structure Identification)**.** *Under Assumptions 2–9, let $\hat{\phi}$ be a representation, possibly learned from the same n samples per environment, with $\hat{\phi} \in \Phi_0$ and $V_{k^*}^*(\hat{\phi}) \leq \eta/2$. Define the estimated structure:*

$$\hat{k}_n := \arg\min_k V_k(\hat{\phi}) \tag{24}$$

Then for the true structure $\mathcal{G}^* = \mathcal{G}_{k^*}$:

$$\mathbb{P}(\hat{k}_n = k^*) \geq 1 - 6\exp\left(-\frac{n\eta^2}{2048\kappa^8}\right) \tag{25}$$

where η is the faithfulness gap (Definition D.2) and κ is the kernel bound (Assumption 2). Proposition 1 gives conditions under which the minimizers of $\mathcal{L}$ satisfy these hypotheses.

Corollary 1 (Identification with Random Features)**.** *Suppose D random features (Definition F.1) achieve uniform kernel error ε_D with probability at least $1-\delta'$, so that $|\widehat{\text{HSIC}}_{\text{RFF}} - \widehat{\text{HSIC}}| \leq C\kappa^2\varepsilon_D$. Under the conditions of Theorem D.2, if $C\kappa^2\varepsilon_D \leq \eta/16$, then*

$$\mathbb{P}(\hat{k}_n^{\text{RFF}} = k^*) \geq 1 - 6\exp\left(-\frac{n\eta^2}{8192\kappa^8}\right) - \delta', \tag{26}$$

which holds for $D = \Omega(\kappa^4\eta^{-2}d_z\log(n/\delta'))$ [36], where d_z is the representation dimension.

The required D grows as η^{-2}: near-equivalent structures need more random features, not only more samples.

Corollary 2 (Adaptive Weight Concentration)**.** *Under the conditions of Theorem D.2, with probability at least $1 - 6\exp(-n\eta^2/(2048\kappa^8))$, the adaptive weights satisfy:*

$$\alpha_{k^*}(\hat{\phi}) \geq 1 - 2\exp(-\beta\eta/4) \tag{27}$$

where η is the faithfulness gap.

Via established structure identification, we now quantify how well the learned representation achieves the corresponding invariance and characterize the minimizers of the objective. Throughout, the loss ℓ is bounded and Lipschitz, so that with probability at least $1 - \delta$ the empirical risks $\hat{\mathcal{R}}_e(w \circ \phi)$ are within $O\big(\sqrt{(\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta))/n}\big)$ of the population risks, uniformly over $\phi \in \Phi$, $w \in \mathcal{W}$ and $e \in \mathcal{E}$, where $\mathcal{C}(\Phi, \mathcal{W})$ is the function class complexity.

Theorem D.3 (Invariance Achievement)**.** *Let ϕ_n minimize $\mathcal{L}(\phi, w)$ with regularization $\lambda > 0$. Under Assumptions 2–9, with probability at least $1 - \delta$:*

$$\min_k V_k(\phi_n) \leq \frac{\epsilon_0}{\lambda} + \frac{\log 3}{\beta} + C_2 n^{-1/2} \tag{28}$$

where ϵ_0 is the tolerance of Assumption 5 and $C_2 > 0$ is a constant depending on λ, κ, δ, m and the function class complexity. Under the conditions of Proposition 1, the minimum is attained at the true structure k^* on the event of that proposition, so the bound holds for $V_{k^*}(\phi_n)$.

The violation decreases at rate $O(n^{-1/2})$ to the level $\epsilon_0/\lambda + \log 3/\beta$, which shrinks as β and λ grow, though a larger λ comes at the cost of prediction accuracy. The following result gives conditions under which the minimizers of $\mathcal{L}$ satisfy the hypotheses of Theorem D.2.

Proposition 1 (Minimizers of the Objective)**.** *Let (ϕ_n, w_n) minimize $\mathcal{L}(\phi, w)$ with n samples per environment. Under Assumptions 2–9, with $\epsilon_0 > 0$ in Assumption 5, let $\lambda \geq 8\epsilon_0/\eta$ and $\beta \geq 2\lambda \log 3/\epsilon_0$. Then there is $n_0 = O\big(\epsilon_0^{-2}(\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta) + \lambda^2\kappa^8 \log(1/\delta))\big)$ such that, for $n \geq n_0$, with probability at least $1 - \delta - 6\exp(-n\eta^2/(2048\kappa^8))$:*

1. $\phi_n \in \Phi_0$ (predictiveness)
2. $V^*_{k^*}(\phi_n) \leq \eta/2$ (near invariance)
3. $\hat{k}_n := \arg\min_k V_k(\phi_n) = k^*$ (structure identification)

We know the objective learns a predictive representation that nearly satisfies the true invariance and identifies the true structure when λ and β are large enough relative to ϵ_0 and η. A larger λ requires more samples through n_0, and taking $\delta = \exp(-cn)$ with a small enough constant $c > 0$ shows that $\mathbb{P}(\hat{k}_n \neq k^*)$ decays exponentially in n.

D.2 Generalization Bounds

We now bound out-of-distribution (OOD) generalization by quantifying how distribution shift affects test performance. To measure the similarity between environments, we introduce a causal distance metric below.

Definition D.3 (Interventional Kernel Distance)**.** For environments $e, e' \in \mathcal{E}$, the interventional kernel distance is:

$$d_{\mathcal{K}}(e, e') := \sup_{S \subseteq T} \|K_S^{(e)} - K_S^{(e')}\|_{\text{op}} \tag{29}$$

where $K_S^{(e)}$ is the causal kernel for environment e (all environments share the sample space $\prod_{v \in V} \Omega_v$) and $\|\cdot\|_{\text{op}}$ is the operator norm on bounded measurable functions: for a signed kernel K, $\|K\|_{\text{op}} := \sup_{\|f\|_\infty \leq 1} \sup_\omega \big|\int f(\omega')K(\omega, d\omega')\big|$. For a test environment, define $d_{\mathcal{K}}(e_{\text{test}}, \mathcal{E}_{\text{train}}) := \min_{e \in \mathcal{E}_{\text{train}}} d_{\mathcal{K}}(e_{\text{test}}, e)$.

This distance captures the maximal change in causal mechanisms across environments. Using this metric, we decompose test error into training error, distribution shift, and statistical estimation error.

Theorem D.4 (Generalization Bound)**.** *Let (ϕ_n, w_n) minimize $\mathcal{L}$ over training environments $\mathcal{E}_{\text{train}}$. For a test environment e_{test}, assume* shift regularity*: for a constant $C_\eta > 0$ and all $\phi \in \Phi$, $w \in \mathcal{W}$ and $e \in \mathcal{E}_{\text{train}}$,*

$$\mathcal{R}_{e_{\text{test}}}(w \circ \phi) \leq \mathcal{R}_e(w \circ \phi) + L_w \cdot d_{\mathcal{K}}(e_{\text{test}}, e) + C_\eta \cdot V^*_{k^*}(\phi). \tag{30}$$

Then, under Assumptions 2–9:

$$\mathcal{R}_{e_{\text{test}}}(w_n \circ \phi_n) \leq \underbrace{\max_{e \in \mathcal{E}_{\text{train}}} \hat{\mathcal{R}}_e(w_n \circ \phi_n)}_{\textit{training error}} + \underbrace{L_w \cdot d_{\mathcal{K}}(e_{\text{test}}, \mathcal{E}_{\text{train}})}_{\textit{shift penalty}} + \underbrace{\xi_n}_{\textit{estimation error}} \tag{31}$$

where the estimation error satisfies:

$$\xi_n = O\left(\sqrt{\frac{\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta)}{n}}\right) + C_\eta \cdot V^*_{k^*}(\phi_n) \tag{32}$$

with $\mathcal{C}(\Phi, \mathcal{W})$ the function class complexity and L_w the Lipschitz constant of Assumption 8. The bound holds with probability at least $1 - \delta$.

The invariance violation $V^*_{k^*}(\phi_n)$ directly contributes to generalization error, motivating the joint optimization of prediction and invariance. Under the conditions of Proposition 1, $\phi_n \in \Phi_0$, so Assumption 7 and Theorem D.3 bound it by $\epsilon_0/\lambda + \log 3/\beta + O(n^{-1/2})$ with high probability.

Corollary 3 (Asymptotic Optimality)**.** *Under the conditions of Theorems D.2–D.4, as $n \to \infty$:*

$$\mathcal{R}_{e_{\text{test}}}(w_n \circ \phi_n) \xrightarrow{p} \inf_{\phi \in \Phi_{k^*}} \mathcal{R}_{e_{\text{test}}}(w^*(\phi) \circ \phi) \tag{33}$$

where $\Phi_{k^*} = \{\phi : \mathcal{I}_{k^*}(\phi)\ holds\}$ is the class of representations satisfying the true invariance.

Theorem D.5 (Sample Complexity)**.** *Under the assumptions of Proposition 1 and Theorem D.4, let $0 < \epsilon \le \epsilon_0$, $\lambda = \max\{8\epsilon_0/\eta, 6C_\eta\epsilon_0/\epsilon\}$ and $\beta \ge 2\lambda \log 3/\epsilon_0$. To achieve structure identification and estimation error $\xi_n \le \epsilon$ jointly with probability $\ge 1-\delta$, the required sample size per environment is:*

$$n = O\left(\frac{\kappa^8}{\eta^2}\log\frac{12}{\delta} + \frac{\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta)}{\epsilon^2}\right) \tag{34}$$

where the second term hides constants depending on κ, C_η and the loss ℓ.

E Proofs

This section contains complete proofs of all theoretical results. We organize proofs by building from foundational kernel independence results to the main structure identification and generalization theorems. We first establish auxiliary concentration results that underpin our main theorems.

Lemma 4 (HSIC Concentration)**.** *Let $(X_i, Y_i)_{i=1}^n$ be i.i.d. samples from P_{XY}, with n a multiple of 4. Under Assumption 2 with kernel bound κ and kernels taking values in $[0, \kappa^2]$ (e.g., Gaussian and delta kernels), the unbiased estimator $\widehat{\text{HSIC}}_u$ defined below satisfies:*

$$\mathbb{P}\left(\left|\widehat{\text{HSIC}}_u(X,Y) - \text{HSIC}(X,Y)\right| > \epsilon\right) \le 2\exp\left(-\frac{n\epsilon^2}{32\kappa^8}\right) \tag{35}$$

for all $\epsilon > 0$, and the estimator of Definition 4.3 satisfies $|\widehat{\text{HSIC}}(X,Y) - \widehat{\text{HSIC}}_u(X,Y)| \le 24\kappa^4/n$.

Proof. Write $k_{ij} := k_{\mathcal{X}}(X_i, X_j)$ and $l_{ij} := k_{\mathcal{Y}}(Y_i, Y_j)$, which lie in $[0, \kappa^2]$. Expanding $\frac{1}{n^2}\text{tr}(K_X H K_Y H)$ gives the V-statistic [14]

$$\widehat{\text{HSIC}}(X,Y) = \frac{1}{n^4}\sum_{i,j,q,r=1}^{n} h_{ijqr}, \qquad h_{ijqr} := \frac{1}{4!}\sum_{(t,u,v,w)} \big(k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{tu}l_{tv}\big), \tag{36}$$

where the inner sum runs over all orderings (t,u,v,w) of (i,j,q,r). Let $\widehat{\text{HSIC}}_u$ be the average of h_{ijqr} over tuples of distinct indices. It is a U-statistic of order 4, and for distinct indices $\mathbb{E}[h_{ijqr}] = \mathbb{E}[k_{\mathcal{X}}(X,X')k_{\mathcal{Y}}(Y,Y')] + \mathbb{E}[k_{\mathcal{X}}(X,X')]\mathbb{E}[k_{\mathcal{Y}}(Y,Y')] - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}[k_{\mathcal{X}}(X,X')]\,\mathbb{E}_{Y'}[k_{\mathcal{Y}}(Y,Y')]\right] = \text{HSIC}(X,Y)$, so $\widehat{\text{HSIC}}_u$ is unbiased. Each term in the sum defining h_{ijqr} lies in $[-2\kappa^4, 2\kappa^4]$, hence so does h_{ijqr}, and Hoeffding's inequality for U-statistics [18] with $n/4$ independent blocks gives

$$\mathbb{P}\left(\left|\widehat{\text{HSIC}}_u - \text{HSIC}\right| > \epsilon\right) \le 2\exp\left(-\frac{2(n/4)\epsilon^2}{(4\kappa^4)^2}\right) = 2\exp\left(-\frac{n\epsilon^2}{32\kappa^8}\right). \tag{37}$$

For the second claim, a fraction $1 - n(n-1)(n-2)(n-3)/n^4 \le 6/n$ of the index tuples has a repeated index, and the averages of h over the two kinds of tuples both lie in $[-2\kappa^4, 2\kappa^4]$, so they differ by at most $4\kappa^4$. Hence $|\widehat{\text{HSIC}} - \widehat{\text{HSIC}}_u| \le 24\kappa^4/n$. □

Lemma 5 (Kernel Regression Consistency)**.** *Under Assumptions 2 and 3, let $\hat{m}(z) = \mathbf{k}_z^\top(K + \gamma I)^{-1}\mathbf{y}$ be the kernel ridge regression estimate for targets bounded by M, and let $\bar{\gamma} := \gamma/n$. Then*

$$\mathbb{E}\left[\|\hat{m} - m^*\|^2_{L^2(P_Z)}\right] \le \frac{\bar{\gamma}}{2}\|m^*\|^2_{\mathcal{H}} + \frac{2\kappa^2 B^2}{n\bar{\gamma}^2}, \tag{38}$$

where $m^*(z) = \mathbb{E}[Y|Z=z]$ is the true conditional expectation and $B := \kappa(M + \kappa\|m^*\|_{\mathcal{H}})$. In particular, the error is $O(n^{-1/3})$ for $\bar{\gamma} = n^{-1/3}$.

Proof. Let $\mathcal{T} := \mathbb{E}[k(Z,\cdot) \otimes k(Z,\cdot)]$ and $\mathcal{T}_n := \frac{1}{n}\sum_{i=1}^n k(z_i,\cdot) \otimes k(z_i,\cdot)$ be the population and empirical covariance operators on $\mathcal{H}$, and let $g := \mathbb{E}[Yk(Z,\cdot)] = \mathcal{T}m^*$ and $g_n := \frac{1}{n}\sum_{i=1}^n y_i k(z_i,\cdot)$. Then $\hat{m} = (\mathcal{T}_n + \bar{\gamma})^{-1} g_n$, and we compare it with $m_{\bar{\gamma}} := (\mathcal{T} + \bar{\gamma})^{-1} g$.

Bias. Let $\{(\lambda_j, e_j)\}$ be the eigenpairs of $\mathcal{T}$ as an operator on $L^2(P_Z)$ and $m^* = \sum_j a_j e_j$, so that $\|m^*\|_{\mathcal{H}}^2 = \sum_j a_j^2/\lambda_j$ by Assumption 3. Since $\lambda_j\bar{\gamma} \le (\lambda_j + \bar{\gamma})^2/4$,

$$\|m_{\bar{\gamma}} - m^*\|_{L^2(P_Z)}^2 = \sum_j \Big(\frac{\bar{\gamma}}{\lambda_j + \bar{\gamma}}\Big)^2 a_j^2 \le \frac{\bar{\gamma}}{4}\|m^*\|_{\mathcal{H}}^2. \tag{39}$$

Variance. Since $\bar{\gamma} m_{\bar{\gamma}} = g - \mathcal{T} m_{\bar{\gamma}}$,

$$(\mathcal{T}_n + \bar{\gamma})(\hat{m} - m_{\bar{\gamma}}) = \frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}[\xi_1], \qquad \xi_i := \big(y_i - m_{\bar{\gamma}}(z_i)\big)\, k(z_i,\cdot), \tag{40}$$

where $\|\xi_i\|_{\mathcal{H}} \le \kappa(M + \kappa\|m_{\bar{\gamma}}\|_{\mathcal{H}}) \le B$ because $\|m_{\bar{\gamma}}\|_{\mathcal{H}} \le \|m^*\|_{\mathcal{H}}$. Since $\|(\mathcal{T}_n + \bar{\gamma})^{-1}\| \le 1/\bar{\gamma}$ and the ξ_i are i.i.d.,

$$\mathbb{E}\|\hat{m} - m_{\bar{\gamma}}\|_{\mathcal{H}}^2 \le \frac{B^2}{n\bar{\gamma}^2}. \tag{41}$$

Combining both parts with $\|f\|_{L^2(P_Z)} \le \kappa\|f\|_{\mathcal{H}}$ and $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ gives the claim. □

E.1 Proof of Lemma 1 (HSIC Characterizes Independence)

Proof. We prove both directions.

(⇐) Independence implies zero HSIC: Assume $X \perp\!\!\!\perp Y$, so $P_{XY} = P_X \otimes P_Y$. The cross-covariance operator is:

$$\mathcal{C}_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)] - \mu_X \otimes \mu_Y \tag{42}$$

where $\phi(x) = k_{\mathcal{X}}(\cdot, x)$ and $\psi(y) = k_{\mathcal{Y}}(\cdot, y)$ are the canonical feature maps.

Under independence:

$$\mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)] = \mathbb{E}_X[\phi(X)] \otimes \mathbb{E}_Y[\psi(Y)] = \mu_X \otimes \mu_Y \tag{43}$$

Therefore $\mathcal{C}_{XY} = 0$, and $\mathrm{HSIC}(X,Y) = \|\mathcal{C}_{XY}\|_{\mathrm{HS}}^2 = 0$.

(⇒) Zero HSIC implies independence: Assume $\mathrm{HSIC}(X,Y) = 0$. By definition:

$$\begin{aligned}\|\mathcal{C}_{XY}\|_{\mathrm{HS}}^2 = \mathbb{E}_{XX'YY'}[k_{\mathcal{X}}(X,X')k_{\mathcal{Y}}(Y,Y')] - 2\mathbb{E}_{XY}\big[\mathbb{E}_{X'}[k_{\mathcal{X}}(X,X')]\,\mathbb{E}_{Y'}[k_{\mathcal{Y}}(Y,Y')]\big] \\ + \mathbb{E}_{XX'}[k_{\mathcal{X}}(X,X')]\mathbb{E}_{YY'}[k_{\mathcal{Y}}(Y,Y')]\end{aligned} \tag{44}$$

where (X,Y) and (X',Y') are independent copies.

This equals the squared MMD between P_{XY} and $P_X \otimes P_Y$ in the product RKHS $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$:

$$\mathrm{HSIC}(X,Y) = \mathrm{MMD}^2(P_{XY}, P_X \otimes P_Y) \tag{45}$$

Since $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ is characteristic (by assumption; e.g., for Gaussian and delta kernels [44]), $\mathrm{MMD}^2 = 0$ implies $P_{XY} = P_X \otimes P_Y$, hence $X \perp\!\!\!\perp Y$. □

E.2 Proof of Lemma 2 (Conditional HSIC Consistency)

Proof. Equivalence. Given $Z = z$, X and Y are ε_X and ε_Y shifted by the constants $m_X(z)$ and $m_Y(z)$, so $X \perp\!\!\!\perp Y \mid Z$ if and only if $\varepsilon_X \perp\!\!\!\perp \varepsilon_Y \mid Z$. Since $(\varepsilon_X, \varepsilon_Y) \perp\!\!\!\perp Z$, the conditional law of $(\varepsilon_X, \varepsilon_Y)$ given Z equals its marginal law, so this is equivalent to $\varepsilon_X \perp\!\!\!\perp \varepsilon_Y$.

Consistency. The population residuals are ε_X and ε_Y. For bounded L-Lipschitz kernels, replacing $\varepsilon_{X,i}$ by $\tilde{x}_i = x_i - \hat{m}_X(z_i)$ changes each Gram entry by at most $L(\|\tilde{x}_i - \varepsilon_{X,i}\| + \|\tilde{x}_j - \varepsilon_{X,j}\|)$. Since centering satisfies $n^{-2}\sum_{i,j}|(HAH)_{ij}| \le 4n^{-2}\sum_{i,j}|A_{ij}|$ and the kernels are bounded,

$$\big|\widehat{\mathrm{HSIC}}(\tilde{X},\tilde{Y}) - \widehat{\mathrm{HSIC}}(\varepsilon_X,\varepsilon_Y)\big| \le \frac{C}{n}\sum_{i=1}^n \big(\|\hat{m}_X(z_i) - m_X(z_i)\| + \|\hat{m}_Y(z_i) - m_Y(z_i)\|\big)$$

for a constant C depending on L and the kernel bounds. By Cauchy–Schwarz and L^2-consistency, the right-hand side converges to 0 in probability. By Lemma 4, $\widehat{\mathrm{HSIC}}(\varepsilon_X, \varepsilon_Y) \xrightarrow{p} \mathrm{HSIC}(\varepsilon_X, \varepsilon_Y)$. Hence $\widehat{\mathrm{HSIC}}(\tilde{X}, \tilde{Y}) \xrightarrow{p} \mathrm{HSIC}(\varepsilon_X, \varepsilon_Y)$, which by Lemma 1 and the equivalence above is 0 if $X \perp\!\!\!\perp Y \mid Z$ and positive otherwise. □

E.3 Proof of Lemma 3 (SoftMin Properties)

Proof. We establish each property.

Part 1: Bounds. For the upper bound, note that for all k:

$$\sum_{j=1}^{3} \exp(-\beta V_j) \geq \exp(-\beta V_k) \tag{46}$$

Taking logarithms and the infimum over k:

$$-\frac{1}{\beta} \log \sum_{j=1}^{3} \exp(-\beta V_j) \leq -\frac{1}{\beta} \log \exp(-\beta \min_k V_k) = \min_k V_k \tag{47}$$

For the lower bound:

$$\sum_{j=1}^{3} \exp(-\beta V_j) \leq 3 \exp(-\beta \min_k V_k) \tag{48}$$

Therefore:

$$-\frac{1}{\beta} \log \sum_{j=1}^{3} \exp(-\beta V_j) \geq -\frac{1}{\beta} \log \left(3 \exp(-\beta \min_k V_k) \right) = \min_k V_k - \frac{\log 3}{\beta} \tag{49}$$

Combining: $\min_k V_k - \frac{\log 3}{\beta} \leq \mathrm{SoftMin}_\beta \leq \min_k V_k$.

Part 2: Limiting behavior. Let $k^* = \arg\min_k V_k$ and assume uniqueness (the general case follows by continuity). Then:

$$\sum_{j=1}^{3} \exp(-\beta V_j) = \exp(-\beta V_{k^*}) \left(1 + \sum_{j \neq k^*} \exp(-\beta (V_j - V_{k^*})) \right) \tag{50}$$

Since $V_j - V_{k^*} > 0$ for $j \neq k^*$:

$$\sum_{j \neq k^*} \exp(-\beta (V_j - V_{k^*})) \to 0 \quad \text{as } \beta \to \infty \tag{51}$$

Therefore:

$$\mathrm{SoftMin}_\beta = -\frac{1}{\beta} \log \exp(-\beta V_{k^*}) - \frac{1}{\beta} \log (1 + o(1)) = V_{k^*} + o(1) \to \min_k V_k \tag{52}$$

Part 3: Gradient. By direct computation:

$$\nabla_\phi \mathrm{SoftMin}_\beta = -\frac{1}{\beta} \cdot \frac{\sum_{k=1}^{3} \exp(-\beta V_k) \cdot (-\beta \nabla_\phi V_k)}{\sum_{j=1}^{3} \exp(-\beta V_j)} \tag{53}$$

$$= \sum_{k=1}^{3} \frac{\exp(-\beta V_k)}{\sum_{j=1}^{3} \exp(-\beta V_j)} \nabla_\phi V_k \tag{54}$$

$$= \sum_{k=1}^{3} \alpha_k(\phi) \nabla_\phi V_k \tag{55}$$

where $\alpha_k(\phi) = \exp(-\beta V_k) / \sum_j \exp(-\beta V_j)$ are the adaptive weights. □

E.4 Proof of Theorem D.1 (Violation Characterization)

Proof. By the Markov property, d-separation in $\mathcal{G}_{k^*}$ implies conditional independence; by faithfulness, d-connection implies dependence. In all three graphs, E enters only through $E \to U$.

$\mathcal{G}_1$. The only parent of $X_z^\perp$ is Y. Every path from E to $X_z^\perp$ either passes through Y as a non-collider or contains an unconditioned collider (Z or X_z); hence $X_z^\perp \perp\!\!\!\perp E \mid Y$ ($\mathcal{I}_1$ holds). The path $E \to U \to Y \to X_z^\perp$ is open without conditioning, so $X_z^\perp \not\perp\!\!\!\perp E$ ($\mathcal{I}_3$ fails); $E \to U \to Y$ is open given $X_z^\perp$, so $Y \not\perp\!\!\!\perp E \mid X_z^\perp$ ($\mathcal{I}_2$ fails).

$\mathcal{G}_2$. Given $X_z^\perp$, the path $E \to U \to X_z^\perp \to Y$ is blocked at $X_z^\perp$, and every other path from E to Y contains an unconditioned collider (X_z or Z); hence $Y \perp\!\!\!\perp E \mid X_z^\perp$ ($\mathcal{I}_2$ holds). The path $E \to U \to X_z^\perp$ is open both unconditionally and given Y, so $\mathcal{I}_3$ and $\mathcal{I}_1$ fail.

$\mathcal{G}_3$. $X_z^\perp$ has no parents, and every path from E to $X_z^\perp$ contains a collider (Y, Z, or X_z); hence $X_z^\perp \perp\!\!\!\perp E$ ($\mathcal{I}_3$ holds). Conditioning on the collider Y opens $E \to U \to Y \leftarrow X_z^\perp$, so $\mathcal{I}_1$ fails; $E \to U \to Y$ is open given $X_z^\perp$, so $\mathcal{I}_2$ fails.

Lemma 1 (for V_3) and Lemma 2 (for V_1, V_2) convert independence into zero violation and dependence into positive violation. Achievability follows because $X_z^\perp \in \Phi_{\epsilon_0} \subseteq \Phi_\epsilon$ (Assumption 5) attains $V_{k^*}^*(X_z^\perp) = 0$ and $V_k^* \geq 0$. Non-achievability follows from Definition D.2, since $\epsilon \leq 2\epsilon_0$ implies $\Phi_\epsilon \subseteq \Phi_0$, and $\eta > 0$ by Assumption 6. □

E.5 Proof of Theorem D.2

Theorem D.2 (Structure Identification)**.** *Under Assumptions 2–9, let $\hat{\phi}$ be a representation, possibly learned from the same n samples per environment, with $\hat{\phi} \in \Phi_0$ and $V_{k^*}^*(\hat{\phi}) \leq \eta/2$. Define the estimated structure:*

$$\hat{k}_n := \arg\min_k V_k(\hat{\phi}) \tag{24}$$

Then for the true structure $\mathcal{G}^* = \mathcal{G}_{k^*}$:

$$\mathbb{P}(\hat{k}_n = k^*) \geq 1 - 6\exp\left(-\frac{n\eta^2}{2048\kappa^8}\right) \tag{25}$$

where η is the faithfulness gap (Definition D.2) and κ is the kernel bound (Assumption 2). Proposition 1 gives conditions under which the minimizers of $\mathcal{L}$ satisfy these hypotheses.

Proof. Apply Assumption 7 with $t = \eta/8$ to each k and take a union bound over $k \in \{1, 2, 3\}$: with probability at least $1 - 3 \cdot 2\exp(-n\eta^2/(64 \cdot 32\kappa^8)) = 1 - 6\exp(-n\eta^2/(2048\kappa^8))$, we have $\sup_{\phi \in \Phi_0} |V_k(\phi) - V_k^*(\phi)| \leq \eta/8$ for all k. The bound is uniform over Φ_0, so it applies to $\hat{\phi} \in \Phi_0$ even when $\hat{\phi}$ is learned from the same samples. On this event, $V_{k^*}(\hat{\phi}) \leq V_{k^*}^*(\hat{\phi}) + \eta/8 \leq 5\eta/8$, while for $k \neq k^*$, Definition D.2 gives $V_k^*(\hat{\phi}) \geq \eta$ and hence $V_k(\hat{\phi}) \geq 7\eta/8$. Therefore $\hat{k}_n = k^*$. Since $\sum_n \exp(-cn) < \infty$ for $c > 0$, the Borel–Cantelli lemma gives $\hat{k}_n \xrightarrow{a.s.} k^*$ when the hypotheses hold for every n. □

E.6 Proof of Corollary 1

Proof. Split the tolerance $\eta/8$ of the proof of Theorem D.2 into two halves. *Sampling error:* Assumption 7 with $t = \eta/16$ and a union bound over k give $\sup_{\phi \in \Phi_0} |V_k(\phi) - V_k^*(\phi)| \leq \eta/16$ for all k with probability at least $1 - 6\exp(-n\eta^2/(256 \cdot 32\kappa^8)) = 1 - 6\exp(-n\eta^2/(8192\kappa^8))$. *Approximation error:* with probability at least $1-\delta'$, $|V_k^{\mathrm{RFF}}(\hat{\phi}) - V_k(\hat{\phi})| \leq C\kappa^2\varepsilon_D \leq \eta/16$. On both events, $V_{k^*}^{\mathrm{RFF}}(\hat{\phi}) \leq \eta/2 + \eta/8 = 5\eta/8$ and $V_k^{\mathrm{RFF}}(\hat{\phi}) \geq \eta - \eta/8 = 7\eta/8$ for $k \neq k^*$, so $\hat{k}_n^{\mathrm{RFF}} = k^*$. By the uniform convergence of random Fourier features [36], $D = \Omega(\varepsilon_D^{-2} d_z \log(n/\delta'))$ features achieve error ε_D with probability $1-\delta'$; setting $\varepsilon_D = \eta/(16C\kappa^2)$ gives $D = \Omega(\kappa^4\eta^{-2} d_z \log(n/\delta'))$. □

E.7 Proof of Corollary 2

Corollary 2 (Adaptive Weight Concentration)**.** *Under the conditions of Theorem D.2, with probability at least* $1 - 6\exp(-n\eta^2/(2048\kappa^8))$*, the adaptive weights satisfy:*

$$\alpha_{k^*}(\hat{\phi}) \geq 1 - 2\exp(-\beta\eta/4) \tag{27}$$

where η *is the faithfulness gap.*

Proof. On the event of the proof of Theorem D.2, $V_{k^*}(\hat{\phi}) \leq 5\eta/8$ and $V_k(\hat{\phi}) \geq 7\eta/8$ for $k \neq k^*$, so $V_k(\hat{\phi}) - V_{k^*}(\hat{\phi}) \geq \eta/4$. Since $\alpha_k \propto \exp(-\beta V_k)$,

$$\alpha_{k^*}(\hat{\phi}) = \frac{1}{1 + \sum_{k \neq k^*} \exp\big(-\beta(V_k(\hat{\phi}) - V_{k^*}(\hat{\phi}))\big)} \geq \frac{1}{1 + 2\exp(-\beta\eta/4)} \geq 1 - 2\exp(-\beta\eta/4), \tag{56}$$

where the last step uses $1/(1+x) \geq 1 - x$ for $x \geq 0$. □

E.8 Proof of Theorem D.3

Theorem D.3 (Invariance Achievement)**.** *Let* ϕ_n *minimize* $\mathcal{L}(\phi, w)$ *with regularization* $\lambda > 0$*. Under Assumptions 2–9, with probability at least* $1 - \delta$*:*

$$\min_k V_k(\phi_n) \leq \frac{\epsilon_0}{\lambda} + \frac{\log 3}{\beta} + C_2 n^{-1/2} \tag{28}$$

where ϵ_0 *is the tolerance of Assumption 5 and* $C_2 > 0$ *is a constant depending on* λ*,* κ*,* δ*,* m *and the function class complexity. Under the conditions of Proposition 1, the minimum is attained at the true structure* k^* *on the event of that proposition, so the bound holds for* $V_{k^*}(\phi_n)$*.*

Proof. Let (ϕ_n, w_n) minimize $\mathcal{L}(\phi, w) = \max_e \hat{\mathcal{R}}_e(w \circ \phi) + \lambda \cdot \mathrm{SoftMin}_\beta(\{V_k(\phi)\})$, which is the objective of Definition 5.4. Write $\hat{\mathcal{R}}(\phi, w) := \max_e \hat{\mathcal{R}}_e(w \circ \phi)$ and $\mathcal{R}(\phi, w) := \max_e \mathcal{R}_e(w \circ \phi)$.

Step 1: Comparison with the invariant component. Let $\phi^\circ := \pi_z^\perp$. By Assumption 5, $\phi^\circ \in \Phi$, and for any $\zeta > 0$ there is $w^\circ \in \mathcal{W}$ with $\mathcal{R}(\phi^\circ, w^\circ) \leq \mathcal{R}^* + \epsilon_0 + \zeta$. Optimality of (ϕ_n, w_n) gives $\mathcal{L}(\phi_n, w_n) \leq \mathcal{L}(\phi^\circ, w^\circ)$. By Lemma 3 Part 1, $\mathrm{SoftMin}_\beta(\{V_k(\phi_n)\}) \geq \min_k V_k(\phi_n) - \log 3/\beta$ and $\mathrm{SoftMin}_\beta(\{V_k(\phi^\circ)\}) \leq V_{k^*}(\phi^\circ)$. Hence

$$\hat{\mathcal{R}}(\phi_n, w_n) + \lambda \min_k V_k(\phi_n) \leq \hat{\mathcal{R}}(\phi^\circ, w^\circ) + \lambda V_{k^*}(\phi^\circ) + \frac{\lambda \log 3}{\beta}. \tag{57}$$

Step 2: Statistical errors. With probability at least $1 - \delta/2$, the empirical risks are uniformly within $\Delta_n = O\big(\sqrt{(\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta))/n}\big)$ of the population risks (Section D.1); since $|\max_e a_e - \max_e b_e| \leq \max_e |a_e - b_e|$, also $|\hat{\mathcal{R}}(\phi, w) - \mathcal{R}(\phi, w)| \leq \Delta_n$ for all $\phi \in \Phi$ and $w \in \mathcal{W}$. On this event, $\hat{\mathcal{R}}(\phi^\circ, w^\circ) \leq \mathcal{R}^* + \epsilon_0 + \zeta + \Delta_n$ and $\hat{\mathcal{R}}(\phi_n, w_n) \geq \mathcal{R}(\phi_n, w_n) - \Delta_n \geq \mathcal{R}^* - \Delta_n$, where the last step uses the definition of $\mathcal{R}^*$. Moreover, Assumption 7 with $t_n := \sqrt{32\kappa^8 \log(4/\delta)/n}$ gives $\sup_{\phi \in \Phi_0} |V_{k^*}(\phi) - V_{k^*}^*(\phi)| \leq t_n$ with probability at least $1 - \delta/2$. Since $\phi^\circ \in \Phi_{\epsilon_0} \subseteq \Phi_0$ and $V_{k^*}^*(\phi^\circ) = 0$ (Theorem D.1), this gives $V_{k^*}(\phi^\circ) \leq t_n$.

Step 3: Conclusion. On both events, which hold together with probability at least $1 - \delta$, (57) gives

$$\min_k V_k(\phi_n) \leq \frac{\epsilon_0 + \zeta + 2\Delta_n}{\lambda} + t_n + \frac{\log 3}{\beta}. \tag{58}$$

Letting $\zeta \to 0$ yields the claim with $C_2 n^{-1/2} := 2\Delta_n/\lambda + t_n$. Under the conditions of Proposition 1, its proof shows that $\hat{k}_n = k^*$ on a subset of these events, so the minimum equals $V_{k^*}(\phi_n)$. □

E.9 Proof of Proposition 1 (Minimizers of the Objective)

Proof. We use the notation and the two events of the proof of Theorem D.3, together with a third event: by Assumption 7 with $t = \eta/8$ and a union bound over k, $\sup_{\phi \in \Phi_0} |V_k(\phi) - V_k^*(\phi)| \leq \eta/8$ for all k

with probability at least $1 - 6\exp(-n\eta^2/(2048\kappa^8))$. The three events hold together with probability at least $1 - \delta - 6\exp(-n\eta^2/(2048\kappa^8))$. Let n_0 be such that $2\Delta_n + \lambda t_n \le \epsilon_0/2$ for $n \ge n_0$; by the rates of Δ_n and t_n, $n_0 = O\big(\epsilon_0^{-2}(\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta) + \lambda^2\kappa^8 \log(1/\delta))\big)$ suffices. The conditions on λ and β give $\epsilon_0/\lambda \le \eta/8$ and $\lambda \log 3/\beta \le \epsilon_0/2$, hence $\log 3/\beta \le \epsilon_0/(2\lambda) \le \eta/16$.

Predictiveness. Empirical violations are nonnegative, so (57) and Step 2 of the proof of Theorem D.3 give

$$\mathcal{R}(\phi_n, w_n) \le \hat{\mathcal{R}}(\phi_n, w_n) + \Delta_n \le \mathcal{R}^* + \epsilon_0 + \zeta + 2\Delta_n + \lambda t_n + \frac{\lambda \log 3}{\beta} \le \mathcal{R}^* + 2\epsilon_0 + \zeta. \quad (59)$$

Letting $\zeta \to 0$ gives $\inf_{w \in \mathcal{W}} \mathcal{R}(\phi_n, w) \le \mathcal{R}^* + 2\epsilon_0$, so $\phi_n \in \Phi_0$.

Structure identification. Since $2\Delta_n/\lambda + t_n \le \epsilon_0/(2\lambda) \le \eta/16$, Theorem D.3 gives

$$\min_k V_k(\phi_n) \le \frac{\epsilon_0}{\lambda} + \frac{\log 3}{\beta} + \frac{2\Delta_n}{\lambda} + t_n \le \frac{\eta}{8} + \frac{\eta}{16} + \frac{\eta}{16} = \frac{\eta}{4}. \quad (60)$$

For $k \ne k^*$, $\phi_n \in \Phi_0$ and Definition D.2 give $V_k^*(\phi_n) \ge \eta$, so $V_k(\phi_n) \ge 7\eta/8 > \eta/4$ on the third event. Hence the minimum is attained only at k^*, that is, $\hat{k}_n = k^*$.

Near invariance. By the previous part, $V_{k^*}(\phi_n) \le \eta/4$, so on the third event $V_{k^*}^*(\phi_n) \le \eta/4 + \eta/8 = 3\eta/8 \le \eta/2$. □

E.10 Proof of Theorem D.4 (Generalization Bound)

Proof. Let $e^* = \arg\min_{e \in \mathcal{E}_{\text{train}}} d_{\mathcal{K}}(e_{\text{test}}, e)$ be the closest training environment, so that $d_{\mathcal{K}}(e_{\text{test}}, e^*) = d_{\mathcal{K}}(e_{\text{test}}, \mathcal{E}_{\text{train}})$. Shift regularity with $\phi = \phi_n$, $w = w_n$ and $e = e^*$ gives

$$\mathcal{R}_{e_{\text{test}}}(w_n \circ \phi_n) \le \mathcal{R}_{e^*}(w_n \circ \phi_n) + L_w \cdot d_{\mathcal{K}}(e_{\text{test}}, \mathcal{E}_{\text{train}}) + C_\eta \cdot V_{k^*}^*(\phi_n). \quad (61)$$

With probability at least $1 - \delta$, the empirical risks are uniformly within $\Delta_n = O\big(\sqrt{(\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta))/n}\big)$ of the population risks (Section D.1), so

$$\mathcal{R}_{e^*}(w_n \circ \phi_n) \le \hat{\mathcal{R}}_{e^*}(w_n \circ \phi_n) + \Delta_n \le \max_{e \in \mathcal{E}_{\text{train}}} \hat{\mathcal{R}}_e(w_n \circ \phi_n) + \Delta_n. \quad (62)$$

Combining the two displays gives the claim with $\xi_n = \Delta_n + C_\eta \cdot V_{k^*}^*(\phi_n)$. □

E.11 Proof of Theorem D.5 (Sample Complexity)

Proof. We derive the sample complexity for joint structure identification and generalization. We run the proofs of Theorem D.3 and Proposition 1 with $\delta/2$ in place of δ, so that the risk deviation Δ_n and the violation deviation t_n hold together with probability at least $1 - \delta/2$.

Requirement 1: Structure identification. The third event of the proof of Proposition 1 fails with probability at most $6\exp(-n\eta^2/(2048\kappa^8)) \le \delta/2$ whenever

$$n \ge \frac{2048\kappa^8}{\eta^2} \log \frac{12}{\delta}. \quad (63)$$

For $n \ge n_0$, the proposition then gives $\phi_n \in \Phi_0$ and $\hat{k}_n = k^*$. Since $\epsilon \le \epsilon_0$ and $\lambda = \max\{8\epsilon_0/\eta, 6C_\eta\epsilon_0/\epsilon\}$ gives $\lambda^2/\epsilon_0^2 \le 64/\eta^2 + 36C_\eta^2/\epsilon^2$, the sample size n_0 of the proposition satisfies

$$n_0 = O\left(\frac{\kappa^8}{\eta^2} \log \frac{1}{\delta} + \frac{\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta) + C_\eta^2\kappa^8 \log(1/\delta)}{\epsilon^2}\right). \quad (64)$$

Requirement 2: Generalization error. From Theorem D.4, $\xi_n = \Delta_n + C_\eta \cdot V_{k^*}^*(\phi_n)$. Since $\phi_n \in \Phi_0$, the violation event gives $V_{k^*}^*(\phi_n) \le V_{k^*}(\phi_n) + t_n$, and since $\hat{k}_n = k^*$, Theorem D.3 bounds $V_{k^*}(\phi_n)$:

$$V_{k^*}^*(\phi_n) \le \frac{\epsilon_0}{\lambda} + \frac{\log 3}{\beta} + \frac{2\Delta_n}{\lambda} + 2t_n. \quad (65)$$

The choice of λ gives $C_\eta \epsilon_0/\lambda \le \epsilon/6$ and, since $\epsilon \le \epsilon_0$, $\lambda \ge 6C_\eta$; the condition $\beta \ge 2\lambda \log 3/\epsilon_0$ gives $C_\eta \log 3/\beta \le C_\eta \epsilon_0/(2\lambda) \le \epsilon/12$. Hence

$$\xi_n \le \frac{\epsilon}{4} + \frac{4}{3}\Delta_n + 2C_\eta t_n, \tag{66}$$

which is at most ϵ once $\Delta_n \le 9\epsilon/32$ and $2C_\eta t_n \le 3\epsilon/8$, that is, for

$$n = O\left(\frac{\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta) + C_\eta^2 \kappa^8 \log(1/\delta)}{\epsilon^2}\right). \tag{67}$$

Combining requirements. Taking the maximum of the three sample sizes and absorbing C_η and κ into the constant of the second term:

$$n = O\left(\frac{\kappa^8}{\eta^2} \log \frac{12}{\delta} + \frac{\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta)}{\epsilon^2}\right) \tag{68}$$

Union bound. The total failure probability is:

$$\mathbb{P}(\hat{k}_n \ne k^* \text{ or } \xi_n > \epsilon) \le \frac{\delta}{2} + \frac{\delta}{2} = \delta \tag{69}$$

Therefore, with probability $\ge 1 - \delta$, both structure identification succeeds and generalization error is $\le \epsilon$. □

E.12 Proof of Proposition 2 (Differentiable CVaR)

Proof. Write $r_e := \hat{\mathcal{R}}_e(w \circ \phi)$ and $K := \rho|\mathcal{E}|$. Recall the CVaR definition:

$$\mathrm{CVaR}_\rho(\{r_e\}) = \inf_{t \in \mathbb{R}} \left\{ t + \frac{1}{K} \sum_{e \in \mathcal{E}} \max(r_e - t, 0) \right\} \tag{70}$$

Convexity. The function $g(t, r) = t + \frac{1}{K}\sum_e \max(r_e - t, 0)$ is jointly convex in (t, r), and minimizing over t preserves convexity, so CVaR_ρ is convex in $r = \{r_e\}$.

Dual form and weights. By linear programming duality,

$$\mathrm{CVaR}_\rho(\{r_e\}) = \max\Big\{ \sum_e p_e r_e : 0 \le p_e \le \frac{1}{K},\ \sum_e p_e = 1 \Big\}. \tag{71}$$

A maximizer p^* fills the largest risks first: $p_e^* = 1/K$ for $r_e > t^*$, $p_e^* = 0$ for $r_e < t^*$, and the environments with $r_e = t^*$ share the remaining weight, where t^* is a minimizer in the definition (a $(1-\rho)$-quantile of $\{r_e\}$). Thus $\sum_e p_e^* = 1$; when K is not an integer, the boundary environment receives a fractional weight, and for $\rho = 1/|\mathcal{E}|$, p^* is the indicator of the worst environment.

Subgradient. The dual form writes $\mathrm{CVaR}_\rho(\{\hat{\mathcal{R}}_e(w \circ \phi)\})$ as a maximum of functions that are differentiable in ϕ, so by Danskin's theorem $\sum_e p_e^* \nabla_\phi \hat{\mathcal{R}}_e$ is a subgradient (in the Clarke sense) w.r.t. ϕ. □

E.13 Proof of Theorem F.1 (Algorithm Convergence)

Proof. Index iterations so that $\theta_{t+1} = \theta_t - s_t g_t$, where θ_t collects the parameters of (ϕ_t, w_t) and g_t is the minibatch gradient of $\mathcal{L}_t := \mathcal{L}(\cdot\,; \beta_t)$ at θ_t, with $\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla \mathcal{L}_t(\theta_t)$ and $\mathbb{E}[\|g_t - \nabla \mathcal{L}_t(\theta_t)\|^2 \mid \mathcal{F}_t] \le \sigma^2$, where $\mathcal{F}_t$ is the filtration up to iteration t.

Step 1: Per-step descent. By L-smoothness and $s_t \le s_0 \le 1/L$,

$$\mathbb{E}[\mathcal{L}_t(\theta_{t+1}) \mid \mathcal{F}_t] \le \mathcal{L}_t(\theta_t) - s_t\Big(1 - \frac{L s_t}{2}\Big)\|\nabla \mathcal{L}_t(\theta_t)\|^2 + \frac{L s_t^2 \sigma^2}{2} \le \mathcal{L}_t(\theta_t) - \frac{s_t}{2}\|\nabla \mathcal{L}_t(\theta_t)\|^2 + \frac{L s_t^2 \sigma^2}{2}. \tag{72}$$

Step 2: Annealing drift. The temperature changes the objective between iterations. Since $\frac{d}{d\beta}\mathrm{SoftMin}_\beta(\{V_k\}) = \mathrm{Ent}(\boldsymbol{\alpha})/\beta^2 \in [0, \log 3/\beta^2]$, for every θ,

$$\mathcal{L}_{t+1}(\theta) - \mathcal{L}_t(\theta) \le \lambda \log 3\Big(\frac{1}{\beta_t} - \frac{1}{\beta_{t+1}}\Big), \tag{73}$$

and these increments sum to at most $\lambda \log 3/\beta_0$.

Step 3: Summing over iterations. Taking expectations in Steps 1 and 2, summing over $t = 1, \dots, T$, and using $\mathcal{L}_t \ge -\lambda \log 3/\beta_0$ (risks are nonnegative, and Lemma 3 Part 1),

$$\frac{1}{2}\sum_{t=1}^{T} s_t\, \mathbb{E}[\|\nabla \mathcal{L}_t(\theta_t)\|^2] \le \mathcal{L}_1(\theta_1) + \frac{2\lambda \log 3}{\beta_0} + \frac{L\sigma^2}{2}\sum_{t=1}^{T} s_t^2. \tag{74}$$

With $s_t = s_0 t^{-1/2}$, $\sum_{t=1}^{T} s_t \ge s_0\sqrt{T}$ and $\sum_{t=1}^{T} s_t^2 \le s_0^2(1 + \log T)$, so

$$\min_{t \le T} \mathbb{E}[\|\nabla \mathcal{L}_t(\theta_t)\|^2] \le \frac{2\mathcal{L}_1(\theta_1) + 4\lambda \log 3/\beta_0 + L\sigma^2 s_0^2(1 + \log T)}{s_0\sqrt{T}} = O\Big(\frac{\log T}{\sqrt{T}}\Big). \tag{75}$$

This establishes Part 1.

Part 2. Apply Corollary 2 with $\hat{\phi} = \phi_T$ and $\beta = \beta_T$.

Part 3. The proofs of Theorem D.3 and Proposition 1 use the optimality of (ϕ_n, w_n) only through $\mathcal{L}(\phi_n, w_n) \le \mathcal{L}(\phi^\circ, w^\circ)$ in Step 1 of the proof of Theorem D.3. For (ϕ_T, w_T) we have $\mathcal{L}(\phi_T, w_T) \le \mathcal{L}(\phi^\circ, w^\circ) + \epsilon_{\mathrm{opt}}$, which adds ϵ_{opt} to the right-hand side of (57). Since $\mathcal{R}(\phi^\circ, w^\circ) \le \mathcal{R}^* + \epsilon_0 + \zeta$, both proofs go through with $\epsilon_0 + \epsilon_{\mathrm{opt}}$ in place of ϵ_0, including in the definitions of Φ_0 and η. □

E.14 Proof of Theorem G.1 (Heterogeneous Structure Identification)

Proof. Fix a feature j. Assumption 7 applied to feature j with $t = \eta^{(j)}/8$ and a union bound over k give, with probability at least $1 - 6\exp(-n(\eta^{(j)})^2/(2048\kappa^8))$, $\sup_{\phi \in \Phi_0^{(j)}} |V_k^{(j)}(\phi) - V_k^{(j)*}(\phi)| \le \eta^{(j)}/8$ for all k.

Features with a canonical structure. Let $\sigma^*(j) = k_j^* \ne \perp$. As in the proof of Theorem D.2, on this event $V_{k_j^*}^{(j)}(\hat{\phi}) \le 5\eta^{(j)}/8$ and $V_k^{(j)}(\hat{\phi}) \ge 7\eta^{(j)}/8$ for $k \ne k_j^*$, so the minimum over k is attained at k_j^*. The feature is then assigned $\perp$ only if $V_{k_j^*}^{(j)}(\hat{\phi}) > \tau_\alpha^{(j)}$, that is, only if the permutation test rejects the true invariance, which by Definition G.2 has probability at most α. A union bound over the two failure events gives the stated bound.

Features without a canonical structure. Let $\sigma^*(j) = \perp$ and $\gamma^{(j)} := \min_k V_k^{(j)*}(\hat{\phi}) > 0$. Under permutation, each $V_k^{(j)}$ is an HSIC estimate between independent variables, which is $O_p(1/n)$ [14], so $\tau_\alpha^{(j)} \to 0$; by Assumption 7, $\min_k V_k^{(j)}(\hat{\phi}) \to \gamma^{(j)}$ in probability. Hence $\mathbb{P}(\min_k V_k^{(j)}(\hat{\phi}) > \tau_\alpha^{(j)}) \to 1$, that is, $\mathbb{P}(\hat{\sigma}_n(j) = \perp) \to 1$.

All features. A union bound of the first part over the d features, using $\eta^{(j)} \ge \eta_{\min}$, gives the simultaneous statement. □

F Algorithm Details

This section presents the complete optimization procedure for structure-agnostic representation learning, including the main algorithm, subroutines, complexity analysis, and convergence guarantees.

F.1 Main Algorithm

Algorithm 1 shows a single run; in practice we run it from $M \ge 10$ initializations and select the final model as described in Section 5.3.

Proposition 2 (Differentiable CVaR)**.** CVaR_ρ *is convex in the environment risks and subdifferentiable. A sub-gradient w.r.t.* ϕ *is* $\partial_\phi \mathrm{CVaR}_\rho = \sum_{e\in\mathcal{E}} p_e^* \nabla_\phi \hat{\mathcal{R}}_e$*,*

where t^* *is the optimal threshold,* $p_e^* = \frac{1}{\rho|\mathcal{E}|}$ *if* $\hat{\mathcal{R}}_e > t^*$*,* $p_e^* = 0$ *if* $\hat{\mathcal{R}}_e < t^*$*, and the environments with* $\hat{\mathcal{R}}_e = t^*$ *share the remaining weight, so that* $\sum_e p_e^* = 1$*.*

The structure-agnostic learning problem admits a nested decomposition:

$$\textbf{Outer:} \quad \boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}\in\Delta^2} \mathcal{L}(\phi^*(\boldsymbol{\alpha}), w^*(\boldsymbol{\alpha}); \boldsymbol{\alpha}) \tag{76}$$

$$\textbf{Inner:} \quad (\phi^*(\boldsymbol{\alpha}), w^*(\boldsymbol{\alpha})) = \arg\min_{\phi,w} \Big(\mathrm{CVaR}_\rho(\{\hat{\mathcal{R}}_e\}) + \lambda \sum_k \alpha_k V_k(\phi)\Big) \tag{77}$$

where $\Delta^2 = \{\boldsymbol{\alpha} \in \mathbb{R}^3_+ : \sum_k \alpha_k = 1\}$ is the probability simplex and $\mathcal{L}(\phi, w; \boldsymbol{\alpha})$ denotes the inner objective.

Algorithm 1 Structure-agnostic Causal Representation Learning (SaCRL)

Input: Multi-environment data $\mathcal{D} = \{D_e\}_{e\in\mathcal{E}}$, regularization λ, initial temperature β_0, final temperature $\beta_{\max}$, learning rate s, iterations T, RFF dimension D, CVaR level ρ
Output: Learned representation ϕ_T, predictor w_T, identified structure $\hat{k}$

// Warm-start with ERM
$(\phi_0, w_0) \leftarrow \arg\min_{\phi,w} \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} \hat{\mathcal{R}}_e(w \circ \phi)$
// Initialize RFF maps
Sample $\{\omega_d, b_d\}_{d=1}^D$ for kernels $k_\mathcal{Z}, k_\mathcal{Y}, k_\mathcal{E}$
for $t = 1, \ldots, T$ **do**
  // Temperature annealing
  $\beta_t \leftarrow \beta_0 + (\beta_{\max} - \beta_0) \cdot \min(2t/T, 1)$
  // Compute representations across all environments
  for each environment $e \in \mathcal{E}$ **do**
    Sample minibatch $\{(x_j^e, y_j^e)\}_{j=1}^B$ from D_e
    $z_j^e \leftarrow \phi_{t-1}(x_j^e)$ for all j
  end for
  Pool data: $\{(z_i, y_i, e_i)\}_{i=1}^N \leftarrow$ all minibatch samples with environment labels
  // Compute structure violations (Algorithm 2)
  $(V_1, V_2, V_3) \leftarrow$ ComputeViolations$(\{(z_i, y_i, e_i)\}_{i=1}^N)$
  // Compute adaptive weights
  $\alpha_k \leftarrow \exp(-\beta_t V_k) / \sum_{j=1}^3 \exp(-\beta_t V_j)$ for $k = 1, 2, 3$
  // Compute distributionally robust prediction loss
  $r_e \leftarrow \frac{1}{B} \sum_{j=1}^B \ell(w_{t-1}(z_j^e), y_j^e)$ for each $e \in \mathcal{E}$
  $\mathcal{L}_{\mathrm{pred}} \leftarrow \mathrm{CVaR}_\rho(\{r_e\}_{e\in\mathcal{E}})$
  // Total loss
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{pred}} + \lambda \cdot \mathrm{SoftMin}_{\beta_t}(\{V_k\}_{k=1}^3)$
  // Gradient update
  $\phi_t \leftarrow \phi_{t-1} - s\nabla_\phi \mathcal{L}$
  $w_t \leftarrow w_{t-1} - s\nabla_w \mathcal{L}$
end for
// Final structure identification
Compute final violations $V_k(\phi_T)$ for $k = 1, 2, 3$
$\hat{k} \leftarrow \arg\min_{k\in\{1,2,3\}} V_k(\phi_T)$
return ϕ_T, w_T, $\hat{k}$

F.2 Subroutines

See Algorithms 2–4.

Definition F.1 (Random Fourier Features)**.** For shift-invariant kernels $k(x, x') = k(x - x')$, the random feature map is $\psi : \mathcal{X} \to \mathbb{R}^D$: $\psi(x) = \sqrt{\frac{2}{D}} \left[\cos(\omega_1^\top x + b_1), \ldots, \cos(\omega_D^\top x + b_D)\right]^\top$, where $\omega_d \sim p(\omega)$ is the spectral density and $b_d \sim \mathrm{Uniform}[0, 2\pi]$.

Algorithm 2 Compute Structure Violations

Input: Pooled data $\{(z_i, y_i, e_i)\}_{i=1}^N$, regularization γ
Output: Violations (V_1, V_2, V_3)
// Structure 1 (anti-causal): Test $\phi(X) \perp\!\!\!\perp E \mid Y$
$\tilde{z} \leftarrow \textsc{KernelResiduals}(\{z_i\}, \{y_i\}, \gamma)$ ▷ Regress Z on Y
$\tilde{e} \leftarrow \textsc{KernelResiduals}(\{e_i\}, \{y_i\}, \gamma)$ ▷ Regress E on Y
$V_1 \leftarrow \widehat{\mathrm{HSIC}}_{\mathrm{RFF}}(\tilde{z}, \tilde{e})$
// Structure 2 (confounded-descendant): Test $Y \perp\!\!\!\perp E \mid \phi(X)$
$\tilde{y} \leftarrow \textsc{KernelResiduals}(\{y_i\}, \{z_i\}, \gamma)$ ▷ Regress Y on Z
$\tilde{e}' \leftarrow \textsc{KernelResiduals}(\{e_i\}, \{z_i\}, \gamma)$ ▷ Regress E on Z
$V_2 \leftarrow \widehat{\mathrm{HSIC}}_{\mathrm{RFF}}(\tilde{y}, \tilde{e}')$
// Structure 3 (confounded-outcome): Test $\phi(X) \perp\!\!\!\perp E$
$V_3 \leftarrow \widehat{\mathrm{HSIC}}_{\mathrm{RFF}}(z, e)$ ▷ No residualization needed
return (V_1, V_2, V_3)

Algorithm 3 Kernel Regression Residuals

Input: Target values $\{a_i\}_{i=1}^n$, conditioning values $\{c_i\}_{i=1}^n$, regularization γ
Output: Residuals $\{\tilde{a}_i\}_{i=1}^n$
Compute kernel matrix: $[K]_{ij} \leftarrow k(c_i, c_j)$
Compute predictions: $\hat{\mathbf{a}} \leftarrow K(K + \gamma I_n)^{-1}\mathbf{a}$ where $\mathbf{a} = [a_1, \ldots, a_n]^\top$
Compute residuals: $\tilde{a}_i \leftarrow a_i - \hat{a}_i$ for $i = 1, \ldots, n$
return $\{\tilde{a}_i\}_{i=1}^n$

F.3 Complexity Analysis

Each iteration of Algorithm 1 has time complexity

$$O\left(|\mathcal{E}| \cdot B \cdot (d_x + d_z) + ND^2 + N^3\right), \tag{78}$$

where $|\mathcal{E}|$ is the number of environments, B the batch size per environment, $N = |\mathcal{E}| \cdot B$ the pooled batch size, D the RFF dimension, and d_x, d_z the input and representation dimensions. The ND^2 term is RFF-HSIC (Algorithm 4); the N^3 term is the kernel ridge solve (Algorithm 3), where one Cholesky factorization per conditioning variable is shared across all regression targets. Both terms depend on the minibatch size N rather than the dataset size, and the total cost is linear in the number of candidate structures.

F.4 Convergence Guarantees

The ERM warm start (ϕ_0, w_0) begins optimization from a predictive representation, away from trivial representations such as a constant ϕ, for which $V_1 = V_3 = 0$.

Algorithm 4 RFF-HSIC Computation

Input: Samples $\{u_i\}_{i=1}^n$, $\{v_i\}_{i=1}^n$, RFF parameters $\{(\omega_d^u, b_d^u)\}_{d=1}^D$, $\{(\omega_d^v, b_d^v)\}_{d=1}^D$
Output: $\widehat{\mathrm{HSIC}}_{\mathrm{RFF}}(U, V)$
// Compute random feature embeddings
for $i = 1, \ldots, n$ **do**
  $\psi_i^u \leftarrow \sqrt{2/D} \cdot [\cos(\omega_1^{u\top} u_i + b_1^u), \ldots, \cos(\omega_D^{u\top} u_i + b_D^u)]^\top$
  $\psi_i^v \leftarrow \sqrt{2/D} \cdot [\cos(\omega_1^{v\top} v_i + b_1^v), \ldots, \cos(\omega_D^{v\top} v_i + b_D^v)]^\top$
end for
// Compute mean embeddings
$\hat{\mu}_u \leftarrow \frac{1}{n}\sum_{i=1}^n \psi_i^u, \quad \hat{\mu}_v \leftarrow \frac{1}{n}\sum_{i=1}^n \psi_i^v$
// Compute cross-covariance
$\hat{\Sigma}_{uv} \leftarrow \frac{1}{n}\sum_{i=1}^n \psi_i^u (\psi_i^v)^\top$
// HSIC estimate
$\widehat{\mathrm{HSIC}} \leftarrow \|\hat{\Sigma}_{uv} - \hat{\mu}_u \hat{\mu}_v^\top\|_F^2$
return $\widehat{\mathrm{HSIC}}$

Theorem F.1 (Algorithm Convergence)**.** *Under Assumptions 2–9, suppose that $\mathcal{L}$ is L-smooth in the parameters of (ϕ, w) for every $\beta \in [\beta_0, \beta_{\max}]$ and that minibatch gradients are unbiased with variance at most σ^2. With step size $s_t = s_0 t^{-1/2}$, $s_0 \leq 1/L$, and linear annealing $\beta_t = \beta_0 + (\beta_{\max} - \beta_0) \cdot \min(2t/T, 1)$, Algorithm 1 satisfies:*

1. ***Stationarity:*** $\min_{t \leq T} \mathbb{E}[\|\nabla \mathcal{L}(\phi_t, w_t)\|^2] \leq O(\log T/\sqrt{T})$
2. ***Weight concentration:*** *if ϕ_T satisfies the conditions of Theorem D.2, then $\alpha_{k^*}^{(T)} \geq 1 - 2\exp(-\beta_T \eta/4)$ with probability at least $1 - 6\exp(-n\eta^2/(2048\kappa^8))$, where k^* is the true structure*
3. ***Approximate minimizers:*** *if $\mathcal{L}(\phi_T, w_T) \leq \min_{\phi,w} \mathcal{L}(\phi, w) + \epsilon_{\text{opt}}$, then Theorem D.3 and Proposition 1 hold for (ϕ_T, w_T) with $\epsilon_0 + \epsilon_{\text{opt}}$ in place of ϵ_0*

where $\mathcal{L}$ at iteration t uses β_t and expectations are over stochastic minibatches.

The proof is in Appendix E. The maximum over environments in $\mathcal{L}$, like CVaR_ρ, is not smooth, so the smoothness condition requires a smoothed version, such as a log-sum-exp over environments. Since $\mathcal{L}$ is non-convex, stationarity does not imply a global minimum; multi-start initialization with selection by the lowest objective value (Section 5.3) targets a small ϵ_{opt}.

G Extension to Heterogeneous Structures

When different feature subsets follow different causal structures, we extend the framework to per-variable structure assignment.

Definition G.1 (Variable-Wise Structure Assignment)**.** Let $X = (X^{(1)}, \ldots, X^{(d)})$ be d-dimensional features. A *heterogeneous causal structure* is a mapping $\sigma^* : \{1, \ldots, d\} \to \{1, 2, 3, \perp\}$ assigning each feature to a canonical structure or $\perp$ (none applies). The variable-wise violation for feature j is $V_k^{(j)}(\phi) := V_k(\phi^{(j)})$, where $\phi^{(j)}$ is the component of ϕ that encodes $X^{(j)}$. The per-feature base class $\Phi_0^{(j)}$ and faithfulness gap $\eta^{(j)}$ are defined as in Appendix D, with $V_k^{(j)}$ in place of V_k.

Definition G.2 (Calibrated Threshold)**.** For significance level $\alpha \in (0, 1)$, the calibrated threshold for feature j is:

$$\tau_\alpha^{(j)} := \max_{k \in \{1,2,3\}} \text{quantile}_{1-\alpha}\left(V_{k,\pi}^{(j)}\right) \tag{79}$$

where $V_{k,\pi}^{(j)}$ is the permutation distribution of $V_k^{(j)}$, obtained by recomputing it with permuted environment labels, under the null hypothesis that $\mathcal{I}_k$ holds. Taking the maximum over k ensures that the violation of the true invariance, whichever it is, exceeds $\tau_\alpha^{(j)}$ with probability at most α when each permutation test has level α.

Definition G.3 (Heterogeneous Objective)**.** The structure-agnostic objective with heterogeneous assignment is:

$$\mathcal{L}_{\text{het}}(\phi, w) := \max_{e \in \mathcal{E}} \hat{\mathcal{R}}_e(w \circ \phi) + \lambda \sum_{j=1}^{d} \text{SoftMin}_\beta\left(\{V_k^{(j)}(\phi)\}_{k=1}^3\right) \tag{80}$$

Theorem G.1 (Heterogeneous Structure Identification)**.** *Under Assumptions 2–9 applied to each feature subset, let $\hat{\phi}$ be a representation, possibly learned from the same n samples per environment, with $\hat{\phi} \in \Phi_0^{(j)}$ and $V_{\sigma^*(j)}^{(j)*}(\hat{\phi}) \leq \eta^{(j)}/2$ for each feature j with $\sigma^*(j) \neq \perp$. Define the estimated per-variable structure:*

$$\hat{\sigma}_n(j) := \begin{cases} \arg\min_{k \in \{1,2,3\}} V_k^{(j)}(\hat{\phi}) & \text{if } \min_k V_k^{(j)}(\hat{\phi}) \leq \tau_\alpha^{(j)} \\ \perp & \text{otherwise} \end{cases} \tag{81}$$

Then for each feature j with true structure $\sigma^*(j) \neq \perp$:

$$\mathbb{P}\left(\hat{\sigma}_n(j) = \sigma^*(j)\right) \geq 1 - 6\exp\left(-\frac{n(\eta^{(j)})^2}{2048\kappa^8}\right) - \alpha \tag{82}$$

where $\eta^{(j)}$ is the faithfulness gap of feature j. For features with $\sigma^*(j) = \perp$, if $\min_k V_k^{(j)*}(\hat{\phi}) > 0$ (no invariance holds for the learned component), then $\mathbb{P}(\hat{\sigma}_n(j) = \perp) \to 1$ as $n \to \infty$. A union

bound over the d features identifies all features with $\sigma^*(j) \neq \perp$ simultaneously with probability at least $1 - 6d\exp(-n\eta_{\min}^2/(2048\kappa^8)) - d\alpha$, where $\eta_{\min} := \min_j \eta^{(j)}$.

Corollary 4 (Heterogeneous Generalization)**.** *Let (ϕ_n, w_n) minimize $\mathcal{L}_{\text{het}}$, and suppose shift regularity (Theorem D.4) holds with $\sum_{j:\sigma^*(j)\neq\perp} V^{(j)*}_{\sigma^*(j)}(\phi)$ in place of $V^*_{k^*}(\phi)$. Then the generalization bound holds with estimation error:*

$$\xi_n^{\text{het}} = O\left(\sqrt{\frac{\mathcal{C}(\Phi, \mathcal{W}) + \log(m/\delta)}{n}}\right) + C_\eta \sum_{j:\sigma^*(j)\neq\perp} V^{(j)*}_{\sigma^*(j)}(\phi_n) \tag{83}$$

The cost of per-feature identification enters through the union bound of Theorem G.1, which adds a $\log d$ term to the sample size.

H Additional Experiments

H.1 Implementation

H.1.1 Hyperparameter Selection

The objective $\mathcal{L}_\rho(\phi, w)$ involves hyperparameters: λ (invariance weight), β (SoftMin temperature), γ (kernel regression regularization), D (RFF dimension), and ρ (CVaR level). Table 9 provides defaults and selection guidance.

Table 9: Recommended hyperparameter settings for Algorithm 1.

Parameter	Default	Range	Selection Strategy
λ	1.0	$[0.5, 5]$	Per experiment, by OOD accuracy (oracle selection)
β_0	1.0	$[0.5, 2]$	Fixed (soft selection initially)
$\beta_{\max}$	50	$[20, 100]$	Higher for sharper selection
γ	$n^{-1/3}$	$[n^{-1/2}, n^{-1/5}]$	Cross-validation for kernel regression
ρ	0.5	$[1/\lvert\mathcal{E}\rvert, 1]$	Smaller ρ when all environments are trusted
D	1000	$[500, 2000]$	Increase if approximation error high

H.1.2 Numerical Stability

- Add small $\epsilon = 10^{-6}$ to diagonal of kernel matrices before inversion
- Clip βV_k to $[-20, 20]$ before exp to prevent overflow
- Normalize HSIC values by their empirical standard deviation for comparable scales

H.1.3 Kernel Selection

We use the following kernels:

- **Representations $k_{\mathcal{Z}}$:** Gaussian RBF with bandwidth $\sigma = \text{median}(\|z_i - z_j\|)$
- **Labels $k_{\mathcal{Y}}$:** Delta kernel for classification; Gaussian RBF for regression
- **Environments $k_{\mathcal{E}}$:** Delta kernel (discrete environments)

H.2 Synthetic DGP Equations

For each canonical structure, we generate $m = 3$ training environments with $d = 10$ features:

- $\mathcal{G}_1$ **(Anti-causal):** $Y \sim \text{Bernoulli}(0.5)$, $Z \sim \mathcal{N}(0, I_5)$, $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I_d)$ where $\sigma_e \in \{0.5, 1.0, 1.5\}$ varies across environments. $X = W_1 \cdot \text{onehot}(Y) + W_2 \cdot Z + \epsilon$ where $W_1 \in \mathbb{R}^{d\times 2}$ and $W_2 \in \mathbb{R}^{d\times 5}$ are fixed random matrices.
- $\mathcal{G}_2$ **(Confounded-descendant):** $U \sim \mathcal{N}(\mu_e, I_3)$ with environment-specific means $\mu_e \in \{(-1, 0, 0), (0, 0, 0), (1, 0, 0)\}$, $X = W_1 \cdot U + \epsilon_X$ where $\epsilon_X \sim \mathcal{N}(0, 0.1 I_d)$, and $Y = \sigma(w_{xy}^\top X + w_{uy}^\top U + \epsilon_Y)$ where $\epsilon_Y \sim \mathcal{N}(0, 0.1)$ and σ is the sigmoid function, with the outcome binarized via Bernoulli sampling.
- $\mathcal{G}_3$ **(Confounded-outcome):** $U \sim \mathcal{N}(0, I_3)$, $X_z^\perp = W_1 \cdot U + \epsilon_1$ where $\epsilon_1 \sim \mathcal{N}(0, \sigma_e^2 I_{d/2})$ with $\sigma_e \in \{0.5, 1.0, 1.5\}$ varying across environments, $Y = \sigma(w^\top X_z^\perp)$ binarized, $X_Z = W_2 \cdot \text{onehot}(Y) + W_3 \cdot U + \epsilon_2$, and $X = (X_Z, X_z^\perp)$.

All weight matrices are sampled from $\mathcal{N}(0, 0.5^2)$ and fixed across runs.

H.3 Real-World Datasets

The PACS dataset [22] is an important benchmark for image classification under domain shift. It contains images from four distinct visual domains (Photo, Art Painting, Cartoon, and Sketch), each depicting the seven object categories (dog, elephant, giraffe, guitar, horse, house, and person) under different styles and appearances. Overall, the dataset includes 9,991 images, with domain-specific subsets ranging from natural photographs to artistic and hand-drawn sketches, making it a challenging testbed for methods to generalize across unseen domains.

H.4 Hyperparameter Sensitivity Results

We sweep the invariance penalty weight λ and SoftMin temperature $\beta_{\max}$ across varying degrees of structural misspecification using the bidirectional DGP from Section 7.6 with $\nu \in \{0.25, 0.50, 0.75\}$. Results are shown in Tables 10 and 11.

Table 10: Sensitivity to invariance penalty weight λ ($\beta_{\max} = 50$ fixed). OOD accuracy (%).

Setting	λ=0.5	λ=1.0	λ=2.0	λ=5.0
ν=0.25 (near-causal)	46.0	46.8	47.8	50.2
ν=0.50 (bidirectional)	69.0	70.3	70.4	73.9
ν=0.75 (near-anticausal)	99.7	99.5	99.5	99.6

Table 11: Sensitivity to SoftMin temperature $\beta_{\max}$ ($\lambda = 5.0$ fixed). OOD accuracy (%).

Setting	$\beta_{\max}$=20	$\beta_{\max}$=50	$\beta_{\max}$=100
ν=0.25 (near-causal)	50.3	50.2	51.3
ν=0.50 (bidirectional)	71.4	73.9	72.9
ν=0.75 (near-anticausal)	99.6	99.6	99.6

λ sensitivity is mild and monotonic in the two harder settings: total variation from λ=0.5 to λ=5.0 is at most $\sim$ 5pp in the hardest settings, with no catastrophic failure. $\beta_{\max}$ is even more stable (variation $\leq$ 2.5pp). When the anti-causal component dominates (ν=0.75), all configurations reach $\sim 99.5\%$. We recommend $\lambda \in [1.0, 5.0]$ and $\beta_{\max} = 50$ as robust defaults.

H.5 Crossover Between Candidate Structures

Table 12 sweeps ν across the crossover between $\mathcal{G}_2$ (ν=0) and $\mathcal{G}_1$ (ν=1) in the bidirectional DGP of Section 7.6, comparing SoftMin with hard $\arg\min_k V_k$ selection. The sweep uses independent seeds; at ν=0.50 it agrees with Table 6 within one standard error. The last column is the permutation p-value of $\min_k V_k$ against the null threshold τ_α (Definition G.2).

Table 12: Adaptive weights and OOD accuracy (%) across the $\mathcal{G}_2$–$\mathcal{G}_1$ crossover.

ν	α_1	α_2	α_3	SoftMin	Hard $\arg\min$	Perm. p
0.40	0.407±0.041	0.489±0.080	0.104±0.058	**59.5**±2.0	56.5±2.7	0.013±0.018
0.45	0.415±0.073	0.478±0.139	0.107±0.075	**66.2**±1.5	62.5±2.9	0.011±0.012
0.50	0.421±0.089	0.422±0.143	0.157±0.071	**72.4**±1.5	68.6±3.6	0.181±0.011
0.55	0.484±0.079	0.417±0.143	0.099±0.077	**79.8**±2.0	76.5±2.2	0.033±0.039
0.60	0.549±0.069	0.346±0.126	0.105±0.077	**87.4**±1.9	84.7±2.2	0.016±0.033

H.6 Margin Calibration

Table 13 bins synthetic runs with known structure by the weight margin $\alpha_{(1)} - \alpha_{(2)}$. Identification accuracy increases monotonically with the margin, so the margin measures how strongly the data support the selected structure.

H.7 Semi-Synthetic Bayesian-Network Benchmarks

Construction. For each benchmark, we select a target Y in a published network, take X as its Markov blanket, and induce environments by soft interventions on a latent parent outside the blanket.

Table 13: Identification accuracy by weight margin on synthetic data.

Margin	Runs	Correct	ID accuracy (%)
< 0.1	48	20	41.7
0.1–0.3	60	39	65.0
0.3–0.5	72	62	86.1
> 0.5	60	57	95.0

The canonical structure is read off the published DAG by the rules of Section 2. $\mathcal{G}_3$ requires two independent root components reconverging at Y, which ALARM (12 roots) supplies twice and ASIA (2 roots) once; $\mathcal{G}_1$ is not derivable under this recipe, since it requires the Markov blanket of Y to consist only of descendants of Y. Before training, each derived structure is verified in simulation: its own invariance holds while the competing one fails. We keep a benchmark only if the competing-invariance violation exceeds 5.0 standard deviations of the permutation null.

Because X is the Markov blanket of Y and the intervened latent lies outside it, $Y \perp\!\!\!\perp E \mid X$ holds in every benchmark, so the identity representation satisfies $\mathcal{I}_2$. The faithfulness gap of the $\mathcal{G}_3$ benchmarks is therefore zero at $\phi(X) = X$, and they test which invariance training selects rather than identifiability in the sense of Theorem D.2.

Departures from the published networks. The SACHS consensus DAG fits its own observational data weakly, so SACHS requires larger interventions to produce contrast. For ASIA, we keep the published structure but re-parameterize two conditional probability tables whose published marginals leave the exogenous branch nearly constant. ALARM requires no adjustment.

Discussion. Five benchmarks span two derivable structures across three networks (three $\mathcal{G}_2$, two $\mathcal{G}_3$), so a method that returns a single structure by default is correct on at most 3/5. alarm_co separates more cleanly than asia_g2 (28.3 vs. 24.6 null standard deviations) but has lower accuracy, because its target has three classes over 12 one-hot columns rather than two. alarm_g2 barely clears the 5.0 threshold and is correspondingly the least reliable. The two $\mathcal{G}_3$ benchmarks come from different networks, so their agreement is not an artifact of a single graph.

H.8 Cross-Fitting Variant

Conditional HSIC via kernel ridge regression residualization can be sensitive to estimation bias. We implement a cross-fitting variant following the SplitKCI protocol [35]: fit kernel regression on fold 1, compute residuals on fold 2, swap and average. Results are displayed in Table 14.

Table 14: Cross-fitting comparison on the bidirectional DGP ($\nu = 0.5$, n=1000 per environment). Independent run of the default configuration.

Variant	OOD Acc (%)
No splitting (default)	73.9 $\pm$ 2.7
Cross-fitted (2-fold)	74.0 $\pm$ 2.7
ERM (reference)	61.2 $\pm$ 4.5

OOD accuracy is virtually identical (73.9 vs. 74.0), confirming that estimation-induced bias from joint residualization does not meaningfully affect downstream performance at the sample sizes used in our experiments. Both variants outperform ERM by $\sim$ 13pp.

H.9 EMA Conditioning Target

Table 15 computes the kernel regressions in V_1 and V_2 on an exponential-moving-average (EMA) copy of the encoder with momentum 0.99 or 0.999, while gradients flow only through the online encoder. Results are nearly unchanged, so SaCRL is robust to this choice.

Table 15: EMA conditioning target on synthetic $\mathcal{G}_2$ (n=2000). Momentum 0 is the default online encoder.

EMA momentum	Stru. ID (%)	OOD Acc (%)
0 (default)	89.0	84.2
0.99	89.4	84.3
0.999	89.8	84.5

H.10 RFF Approximation Error

Table 16 reports the mean absolute difference between RFF-based and exact HSIC on synthetic data (n=2000). The error decreases steadily with D, from 1.42×10^{-3} at D=100 to 4.4×10^{-4} at D=2000, consistent with the η^{-2} scaling of the required D in Corollary 1.

Table 16: RFF approximation error versus exact HSIC (synthetic, n=2000).

D	100	250	500	1000	2000
Mean $\lvert\widehat{\mathrm{HSIC}}_{\mathrm{RFF}} - \widehat{\mathrm{HSIC}}\rvert$ ($\times 10^{-4}$)	14.2	12.9	8.4	7.7	4.4

H.11 Effect of the Number of Training Environments

Table 17 reports the effect of varying the number of training environments m on synthetic $\mathcal{G}_2$ data ($n = 2000$). With only two environments, structure identification is unstable due to insufficient contrast for testing the three invariances (with two environments, each column of the diversity matrix $\mathbf{D}$ (Definition C.18) is determined by the single difference between them). Using three or more environments yields reliable identification ($\geq 89\%$) and OOD accuracy ($\geq 84\%$), with diminishing returns beyond $m = 4$.

Table 17: Effect of number of training environments m on synthetic $\mathcal{G}_2$ data (n=2000).

m	Stru. ID. Acc (%)	OOD Acc (%)
2	71.5 $\pm$4.2	74.3 $\pm$2.8
3	89.0 $\pm$1.6	84.2 $\pm$1.3
4	96.5 $\pm$1.2	85.1 $\pm$1.1
5	97.0 $\pm$0.9	85.4 $\pm$0.9

H.12 Computational Efficiency and Runtime Analysis

H.12.1 Training Time and Memory

Table 18 compares training time and memory. With D=1000 RFF features, our method adds modest overhead ($\sim 30\%$) over ERM while avoiding the $O(n^2)$ complexity of exact HSIC. For Colored MNIST (10K images), exact kernel matrices would require > 10 GB GPU memory and $4\times$ training time; RFF introduces negligible accuracy loss while enabling scalability.

H.12.2 Per-Epoch Runtime Breakdown

Table 19 breaks down the training time into different components. We observe that computing HSIC via random Fourier features (RFF) significantly reduces the computational cost (Table 18).

H.13 Fine-Tuned Backbone

Fine-tuning ResNet18 end-to-end (PACS, Photo held out, 3 seeds) raises all methods by 1.0–2.0pp; SaCRL (95.0) is on par with ERM (95.1) and IRM (95.2) (Table 20). Photo is ImageNet-like, so the pretrained backbone is near saturation and the choice of objective matters little. With the frozen backbone, SaCRL's gains concentrate on the stylized domains and on VLCS (Table 5); we did not fine-tune on those domains.

Table 18: Training time and memory for 100 epochs on Colored MNIST (10K samples).

Method	Time (min)	GPU Memory	Relative
ERM	8.3	2.1 GB	1.0×
IRM	9.8	2.4 GB	1.18×
SaCRL (RFF, D=1000)	10.9	2.7 GB	1.31×
SaCRL (exact HSIC)	33.2	11.8 GB	4.0×

Table 19: Runtime breakdown per epoch on PACS.

Component	Time (s)	Fraction
Forward pass	4.59	42%
HSIC computation (RFF)	3.02	28%
Kernel regression	1.75	16%
Backward pass	1.19	11%
Other	0.34	3%
Total	10.9	100%

Table 20: PACS (Photo held out): frozen vs. fine-tuned ResNet18. OOD accuracy (%), 3 seeds.

Method	Frozen	Fine-tuned	Δ
ERM	93.2	95.1	+1.9
IRM	93.2	95.2	+2.0
VREx	93.0	94.9	+1.9
CIRCE	92.5	94.1	+1.6
SaCRL	94.0	95.0	+1.0

I Limitation

Our framework targets three canonical structures arising under partial observability and confounding-mediated environment effects; *cyclic systems, multi-confounder dynamics, or settings where E directly rewires X–Y fall outside this scope*, though the heterogeneous extension and the empirical robustness study (Section 7.6) show graceful degradation under misspecification. In such cases, SaCRL signals the mismatch: if $\min_k V_k$ exceeds the permutation threshold τ_α (Definition G.2), the data do not support a canonical structure, and the selected structure should not be interpreted as identified. Structure identification further requires sufficient environment diversity (Assumption 1) and at least $m \geq 3$ training environments (Appendix H.11); under near-identical environments, the invariance penalty loses informativeness even though identification remains stable. Conditional HSIC estimation depends on kernel regression quality and may benefit from cross-fitting in low-sample regimes, per results in Table 14.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state our three contributions (unified measure-theoretic framework, efficient algorithm with linear scaling in candidate structures, and theoretical guarantees), all of which are supported by Section 5 and Section 6 (theory) and Section 7 (experiments).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed at the end of Section 8, including the scope of three canonical structures, the requirement of sufficient environment diversity (Assumption 1), and sensitivity of conditional HSIC estimation; the empirical robustness study in Section 7.6 characterizes graceful degradation under misspecification.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are stated in Appendix C.6; theorems and lemmas are numbered and cross-referenced throughout Section 6; complete proofs are provided in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Algorithm 1 and its subroutines are fully described in Appendix F; full hyperparameters, kernel choices, and numerical-stability practices are in Appendix H.1; synthetic data-generating equations are in Appendix H.2; code is released anonymously (link in abstract).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is released at the anonymized URL in the abstract. All datasets used (Colored MNIST, PACS, VLCS, OfficeHome) are publicly available standard benchmarks with citations provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 7.1 specifies architectures, optimizer, learning rates, environment sampling, and main hyperparameters (λ, β schedule, ρ, RFF dimension); full hyperparameter table and selection strategy are in Appendix H.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All main results report standard errors over 5 random seeds (20 seeds for Table 1). Error bars represent the standard error of the mean across seeds; the random factor is the seed controlling initialization, mini-batch order, and data subsampling.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources, per-experiment training time, GPU memory usage, and per-epoch runtime breakdown are reported in Appendix H.12.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The work is foundational (representation learning theory and method) using standard public benchmarks; no human subjects, sensitive data, or deployment-stage risks are involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is foundational research on causal representation learning theory and methodology. The contributions (a learning algorithm with theoretical guarantees) are not tied to a specific application or deployment, and we do not foresee a direct path to negative societal impact beyond general considerations applicable to any robust ML method.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: All experiments use existing public benchmarks, and the released code is a standard research implementation that poses no additional misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets used (MNIST, PACS, VLCS, OfficeHome) are properly cited and used under their original research licenses. PACS [22], VLCS [12], and OfficeHome [48] are publicly available for non-commercial research use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release our SaCRL implementation as a new asset at the anonymized URL in the abstract, with documentation and reproduction instructions in the repository README.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human-subjects research, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as part of the core methodology, scientific contribution, or experimental design. Any usage was limited to writing/editing assistance, which per NeurIPS policy does not require declaration.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.